

I limiti strutturali del ragionamento statistico: causalità, impossibilità e Intelligenza Artificiale

Introduzione: l'asimmetria che non svanisce

Vi sono asimmetrie che la matematica rivela non come accidenti del mondo bensì come necessità logiche iscritte nella struttura medesima dei problemi. L'asimmetria tra statistica e causalità appartiene a questa categoria: non è una difficoltà tecnica che raffinamenti metodologici potranno colmare, né una lacuna di dati che campioni più ampi potranno sanare. È un confine di principio, dimostrabile con la stessa necessità con cui si dimostra l'incompletezza dei sistemi formali sufficientemente ricchi o l'indcidibilità del problema della terminazione. Come quei risultati, anche questo confine non chiude un programma di ricerca: lo orienta, indicando dove non ha senso cercare e dove invece il territorio rimane aperto.

Questo giudizio richiede, prima ancora di qualunque argomentazione tecnica, una precisazione sul tipo di necessità invocata. La geometria euclidea è necessaria in senso logico-deduttivo: le sue proposizioni seguono dagli assiomi per pura catena sintattica, e la negazione di un teorema è una contraddizione. I teoremi di Gödel e Turing sono necessari in senso logico-computazionale: la loro forza deriva dalla struttura degli enunciati autoreferenziali e dalla nozione di computabilità, indipendentemente dall'implementazione fisica di qualunque macchina. La strettezza della gerarchia causale — il fatto che nessun algoritmo possa derivare informazione interventistica dalla sola distribuzione osservazionale — è necessaria in senso logico-strutturale: dipende dalla definizione stessa di intervento come operazione che recide archi causali e modifica il meccanismo generatore dei dati, definizione che i dati osservazionali non possono replicare perché ne sono prodotto, non origine.

Questa precisazione non è pedanteria: è essenziale per valutare le affermazioni ricorrenti nella letteratura sull'intelligenza artificiale come nel dibattito più largo sulla scienza dei dati, secondo cui il confine tra correlazione e causalità sarebbe essenzialmente un problema di campione insufficiente, di confonditori non inclusi, di modelli mal specificati. La risposta che questi teoremi impongono è più radicale: il confine esiste anche nel limite ideale di campioni infiniti, di modelli perfettamente specificati nella loro componente osservazionale, di algoritmi senza errore computazionale. È una distanza tra *tipi* di conoscenza, non tra *quantità* di conoscenza.

La genealogia filosofica del problema

Il problema dell'inferenza causale dalle osservazioni ha una storia che precede di secoli la formalizzazione matematica che ne ha permesso la risoluzione — parziale — nel Novecento.

Aristotele distingueva quattro specie di causa — materiale, formale, efficiente e finale — e riconosceva nella causa efficiente, quella che noi oggi chiameremmo causazione lineare o meccanicistica, il tipo epistemicamente più accessibile ma non l'unico rilevante per la comprensione del mondo naturale. La scolastica medievale elaborò questa distinzione attraverso le nozioni di causa prossima e causa remota, di causa per se e causa per accidens, di causa necessaria e causa sufficiente. Questi strumenti concettuali, per quanto privi di un apparato formale, anticipano in modo notevole le distinzioni che la moderna teoria causale ha reso precise.

L'empirismo britannico segnò una svolta decisiva, anche se in direzione apparentemente scettica. David Hume, nella *Ricerca sull'intelletto umano* del 1748 e ancor più nel *Trattato della natura umana* del 1739, argomentò con penetrante chiarezza che la relazione causale non è mai osservabile direttamente: ciò che l'esperienza fornisce è la congiunzione costante di eventi — the constant conjunction — e l'inferenza alla connessione necessaria è un'abitudine della mente, una proiezione psicologica che non ha corrispondente nel mondo esterno. L'argomentazione humeana è epistemicamente acuta e ontologicamente deflazionista: essa non nega che vi siano relazioni causali nel mondo, ma sostiene che la nostra conoscenza di esse è irriducibilmente indiretta e si fonda su attese induttive piuttosto che su percezioni dirette.

La formulazione humeana, pur nella sua acutezza, non fornisce strumenti per distinguere operativamente tra congiunzione causale e congiunzione spuria — tra la correlazione che riflette un meccanismo e quella che è artefatto di selezione, di confondimento, di coincidenza. John Stuart Mill tentò di colmare questa lacuna con i suoi *Metodi di scoperta causale* — il metodo della concordanza, della differenza, dei residui, della variazione concomitante — ma si trattava di euristiche procedurali che non risolvevano il problema di principio: tutti i metodi di Mill presuppongono che le variabili rilevanti siano già state identificate e osservate, e che il confondimento sia stato già escluso per via sperimentale o teorica.

La tradizione macroeconomica e econometrica del Novecento — da Trygve Haavelmo alla Commissione Cowles, dai modelli di equazioni strutturali alle variabili strumentali — affronta il problema causale in modo più formale ma spesso senza esplicitarlo come tale. Haavelmo, nella sua tesi del 1944 *The Probability Approach in Econometrics*, comprese che i parametri delle equazioni strutturali di un sistema economico non sono meri coefficienti di regressione ma descrivono meccanismi causali autonomi — nel senso che rimangono stabili quando si interviene sul sistema. Questa intuizione, sviluppata in seguito da Christopher Sims con i modelli VAR e da David Hendry con la teoria della super-esogeneità, anticipa in modo sorprendente il principio di autonomia causale che Pearl renderà esplicito nella teoria degli SCM. Tuttavia nessuno di questi contributi fornisce un calcolo formale che permetta di decidere algebricamente quando una quantità causale è identificabile e quando non lo è: questa è la lacuna che il do-calculus di Pearl colmerà.

Il principio della causa comune e la genealogia reichenbachiana

Tra i precursori filosofici della teoria causale formale, la figura di Hans Reichenbach occupa un posto di primo piano che la letteratura tecnica tende a trascurare. Nel suo postumo *The Direction of Time* (1956) e nel precedente *The Theory of Probability* (1949), Reichenbach enunciò il principio che porta il suo nome: se due eventi A e B sono statisticamente correlati e nessuno dei due causa l'altro, allora esiste una causa comune C che precede entrambi

temporalmente e che, condizionato su C, rende A e B statisticamente indipendenti. Formalmente:

$A \perp B \mid C$, con C antecedente causale comune di A e B.

Questo principio — il Common Cause Principle di Reichenbach — è il nucleo filosofico della struttura di d-separazione che Verma e Pearl formalizzeranno decenni dopo. La condizione di Markov causale, per cui ogni variabile è indipendente dai suoi non-discendenti condizionata sui propri genitori, è la generalizzazione algebrica del principio reichenbachiano all'intero DAG: le dipendenze residue dopo il condizionamento sui genitori sono esattamente le dipendenze non spiegate da cause comuni esplicite nel modello. Quando un modello viola la condizione di Markov, lo fa perché è incompletamente specificato — mancano cause comuni che, se incluse, ristabilirebbero le indipendenze attese. Il confondimento latente è la manifestazione tecnica del fallimento del principio reichenbachiano rispetto alle variabili effettivamente misurate.

La direzione temporale che Reichenbach introduce — la causa precede l'effetto — è un elemento che la teoria causale moderna di Pearl tratta in modo esplicito ma pragmatico: il tempo non è sempre presente nei modelli causali strutturali, ma quando lo è, introduce un vincolo parziale sulle possibili orientazioni degli archi del DAG che riduce l'ambiguità della classe di equivalenza di Markov. Nelle serie temporali stazionarie, la precedenza temporale è l'unica fonte di identificabilità direzionale disponibile senza esperimenti di intervento — una questione che la Parte Quarta affronterà nel contesto dei modelli Granger-causali e delle loro limitazioni rispetto alla causalità interventistica.

Il contributo formale di Wright e la svolta del Novecento

La prima formulazione tecnica del problema dell'identificazione causale appartiene a Sewall Wright, il genetista che negli anni Venti sviluppò il metodo dell'analisi dei percorsi (*path analysis*) per studiare la trasmissione dei caratteri ereditari. Wright comprese che le correlazioni tra caratteri di organismi correlati geneticamente non si lasciano interpretare senza assumere esplicitamente una struttura delle dipendenze causali tra i caratteri. Introdusse i *diagrammi di percorso* come rappresentazioni della struttura causale assunta, e derivò da essi le equazioni che legano le correlazioni osservate ai parametri causali — le equazioni di Wright, nella loro forma più semplice:

$$r_{\{XY\}} = \sum_{\{p \in \text{percorsi}(X \rightarrow Y)\}} \prod_{\{\text{archi} \in p\}} c_{\text{arco}}$$

dove la somma è estesa a tutti i percorsi causali — diretti e attraverso variabili intermedie — che connettono X a Y nel diagramma, e c_{arco} denota il coefficiente di percorso dell'arco corrispondente.

Il contributo di Wright è metodologicamente fondamentale per tre ragioni distinte. La prima: introduce la distinzione tra struttura causale qualitativa (il grafo dei percorsi) e parametri causali quantitativi (i coefficienti di percorso), riconoscendo che l'una non è riducibile all'altra. La seconda: esplicita come le correlazioni osservate dipendano sia dalla struttura qualitativa sia dai parametri quantitativi, fornendo le equazioni di identificazione. La terza: riconosce che alcune direzioni causali non sono identificabili dalle sole correlazioni, mentre

altre lo sono — un'anticipazione della teoria delle classi di equivalenza che verrà sviluppata settant'anni dopo.

Il passaggio successivo cruciale è il riconoscimento — operato da molti economisti ma sistematizzato da Working nel 1927 e da Koopmans nel 1949 — del problema dell'identificazione come questione formale distinta dal problema della stima. Un sistema di equazioni strutturali può essere sotto-identificato (infiniti valori dei parametri compatibili con la distribuzione osservazionale), esattamente identificato (un unico valore compatibile) o sovra-identificato (restrizioni verificabili sui parametri). Questa tripartizione, sviluppata nel contesto dei modelli econometrici lineari, è la precursore della nozione moderna di identificabilità causale che la Parte Seconda formalizzerà nella sua versione più generale.

Il framework dei risultati potenziali di Rubin e la sua relazione con il formalismo di Pearl

Un'assenza significativa nelle trattazioni tecniche del problema causale è il confronto sistematico tra il framework degli SCM di Pearl e la tradizione statistica degli *potential outcomes* (risultati potenziali), sviluppata da Jerzy Neyman nel 1923 per l'analisi di esperimenti agrari randomizzati e sistematicamente estesa da Donald Rubin a partire dal 1974 in una serie di lavori fondamentali. Questa tradizione — il modello di Neyman-Rubin, o modello dei risultati potenziali, spesso indicato con la sigla PO (Potential Outcomes) o come SUTVA framework — è dominante in econometria, in statistica applicata e nell'analisi delle politiche pubbliche, ed è distinta nel formalismo, nella terminologia e nelle enfasi rispetto al framework di Pearl, pur essendo ad esso matematicamente equivalente al livello-2 della gerarchia causale.

Il punto di partenza del framework PO è la definizione dell'effetto causale come confronto tra risultati potenziali. Per ogni unità sperimentale i e per ogni possibile valore del trattamento t appartenente all'insieme dei trattamenti $\mathcal{T}$, si definisce $Y_i(t)$ come il risultato potenziale dell'unità i se ricevesse il trattamento t . Questa quantità è controfattuale nel senso preciso: se l'unità i riceve il trattamento t^0 , allora $Y_i(t)$ per $t \neq t^0$ è un esito che non si osserverà mai per quell'unità. L'effetto causale individuale del trattamento t rispetto al trattamento t' per l'unità i è la differenza $Y_i(t) - Y_i(t')$, e l'effetto medio del trattamento (ATE, Average Treatment Effect) è:

$$ATE = \mathbb{E}[Y_i(t) - Y_i(t')] = \mathbb{E}[Y_i(t)] - \mathbb{E}[Y_i(t')]$$

L'assunzione fondamentale del framework PO è la SUTVA (Stable Unit Treatment Value Assumption, Rubin 1980): (a) il risultato potenziale di ogni unità non dipende dal trattamento assegnato ad altre unità (assenza di interferenza tra unità); (b) per ogni unità e ogni trattamento, esiste un'unica versione del risultato potenziale, indipendentemente dal meccanismo attraverso cui il trattamento è stato somministrato. La SUTVA è l'equivalente funzionale dell'assunzione di modularità degli SCM di Pearl: ogni unità ha equazioni strutturali autonome che producono risultati determinati dal proprio trattamento, indipendentemente dal trattamento delle altre.

La connessione algebrica tra i due framework è stabilita dal seguente teorema di equivalenza, che Pearl formalizza in *Causality* (2000, Cap. 7) e che Richardson e Robins (2013) analizzano in dettaglio nel contesto delle Single World Intervention Graphs (SWIG):

Teorema di equivalenza PO-SCM. Sia $M = \langle U, V, \mathcal{F}, P(U) \rangle$ un modello causale strutturale con grafo aciclico G , e sia $Y_x(u)$ il controfattuale di Y per l'unità u sotto l'intervento $do(X=x)$. Allora $Y_x(u)$ coincide con il risultato potenziale $Y(x)$ dell'unità u nel framework PO, per ogni valore di u e x . La distribuzione $P(Y_x)$ del controfattuale coincide con $P(Y(x))$ della distribuzione del risultato potenziale nella popolazione, per ogni distribuzione di U . Le due formalizzazioni sono quindi coestensive al livello-3 della gerarchia causale.

La differenza tra i due framework non è ontologica ma metodologica e di enfasi. Il framework PO privilegia la manipolazione sperimentale come via maestra all'inferenza causale: l'assegnazione casuale del trattamento garantisce che $Y(t)$ sia indipendente da T per ogni t , rendendo $\mathbb{E}[Y \mid T=t] = \mathbb{E}[Y(t)]$ — l'uguaglianza tra la distribuzione osservazionale condizionale e la distribuzione del risultato potenziale. Questa uguaglianza è la base teorica del trial randomizzato come gold standard dell'inferenza causale, e la sua violazione è il problema del confondimento nel linguaggio PO, equivalente al problema del bias di confondimento nel linguaggio SCM.

Il framework SCM privilegia invece la struttura del meccanismo generatore dei dati: il grafo causale rende esplicite le dipendenze strutturali, il do-calculus fornisce le condizioni di identificabilità per inferenza causale da dati osservazionali, e la teoria delle classi di equivalenza di Markov quantifica l'ambiguità irriducibile. I due framework si integrano reciprocamente: il PO è potente per l'analisi degli esperimenti e per la comunicazione delle assunzioni di identificabilità in termini di ignorability, mentre gli SCM sono potenti per la modellazione strutturale, la scoperta causale e la dimostrazione dei teoremi di impossibilità che costituiscono il nucleo di questa monografia.

Un elemento cruciale del framework PO, non sempre presente nelle trattazioni degli SCM, è la distinzione tra diversi aestimandi causali: l'ATE, il LATE (Local Average Treatment Effect, Imbens e Angrist 1994, che stima l'effetto causale per i *complier* in presenza di variabili strumentali), l'ATT (Average Treatment Effect on the Treated) e il CATE (Conditional ATE, l'effetto eterogeneamente variabile in funzione delle caratteristiche dell'unità). Questa tipologia di stimandi ha un corrispondente esatto nel framework SCM — l'ATE è $\mathbb{E}[Y_{\{X=1\}} - Y_{\{X=0\}}]$, l'ATT è $\mathbb{E}[Y_{\{X=1\}} - Y_{\{X=0\}} \mid X=1]$, il CATE è $\mathbb{E}[Y_{\{X=1\}} - Y_{\{X=0\}} \mid Z=z]$ — ma la mappa tra i due sistemi di notazione è un ostacolo pratico alla comunicazione interdisciplinare che questa monografia intende contribuire a dissipare.

I metodi quasi-sperimentali come approssimazioni degli interventi

La distinzione tra dati osservazionali e dati interventistici — centrale in tutta la monografia — rischia di apparire come una dicotomia netta tra esperimento randomizzato (perfettamente interventistico) e osservazione passiva (puramente associativa), mentre la pratica scientifica conosce una gamma continua di disegni di ricerca che si collocano tra questi due estremi. I metodi quasi-sperimentali — la regressione con discontinuità (RDD), le differenze-nelle-differenze (DiD), il controllo sintetico, le variabili strumentali (IV) — sono approcci che approssimano la struttura interventistica sfruttando variazioni naturali nel meccanismo di assegnazione al trattamento. La loro analisi formale nel linguaggio degli SCM rivela sia la loro forza relativa rispetto alla pura osservazione sia i residui di ambiguità che li separano dall'esperimento ideale.

Il metodo delle variabili strumentali è quello che ammette la formulazione più precisa nel linguaggio della teoria causale. Uno strumento Z per la relazione causale $X \rightarrow Y$ è una variabile che soddisfa tre condizioni causali: (i) Z causa X (rilevanza); (ii) Z non causa Y direttamente, se non attraverso X (esclusione); (iii) Z è indipendente dalle variabili di confondimento U (esogeneità). Nel grafo causale, queste condizioni richiedono che Z sia un genitore diretto di X , che non vi sia percorso diretto da Z a Y che aggiri X , e che non vi siano archi bidirezionali tra Z e Y o tra Z e U . L'estimatore IV sotto linearità — $\beta_{IV} = \text{Cov}(Z, Y) / \text{Cov}(Z, X)$ — identifica il coefficiente causale β di $X \rightarrow Y$ al netto del confondimento U , e la sua derivazione nel linguaggio degli SCM segue direttamente dalla formula front-door generalizzata. Sotto linearità:

$$\mathbb{E}[Y \mid \text{do}(X=x)] = \beta \cdot x, \text{ con } \beta = \text{Cov}(Z, Y) / \text{Cov}(Z, X)$$

La condizione di esclusione — che Z non influenzi Y attraverso percorsi che bypassino X — è l'assunzione critica non testabile della IV, e corrisponde esattamente all'assunzione che Z non sia presente come genitore diretto di Y nel grafo causale. La sua violazione — lo strumento debole o impuro — è la forma più comune di fallimento del metodo IV, ed è un esempio paradigmatico di come le assunzioni qualitative sulla struttura causale del grafo siano ineliminabili anche nei metodi più sofisticati dell'econometria.

La discontinuità di regressione (RDD) sfrutta il fatto che l'assegnazione al trattamento dipende in modo discontinuo da una variabile di forza (*running variable*) attorno a una soglia critica. Se l'unità è trattata quando la running variable V supera la soglia c — $T = \mathbf{1}[V \geq c]$ — e se non vi è manipolazione strategica attorno alla soglia, allora la distribuzione delle variabili di confondimento è continua attorno a c per costruzione, mentre il trattamento salta discontinuamente. Questo crea una situazione localmente equivalente a un esperimento randomizzato: le unità con V appena sopra c e appena sotto c sono, in media, equivalenti tranne per il trattamento ricevuto. L'effetto causale stimato è il LATE al valore di forza c :

$$\mathbb{E}[Y(1) - Y(0) \mid V = c] = \lim_{v \rightarrow c+} \mathbb{E}[Y \mid V=v] - \lim_{v \rightarrow c-} \mathbb{E}[Y \mid V=v]$$

Nel linguaggio degli SCM, l'RDD sfrutta l'assunzione che il meccanismo che genera la running variable V sia esogeno rispetto al confounder U nel vicinato della soglia — un'assunzione locale di ignorabilità che i dati non possono verificare direttamente ma che può essere parzialmente saggiata con test di densità (McCrary test) e test di continuità delle covariate predeterminate.

Il metodo delle differenze-nelle-differenze (DiD) assume che, in assenza del trattamento, il gruppo trattato avrebbe seguito la stessa tendenza temporale del gruppo di controllo (assunzione delle tendenze parallele). Nel linguaggio causale, questa ipotesi equivale a richiedere che la distribuzione del confounder latente U sia la stessa nel tempo per entrambi i gruppi — un'assunzione di stazionarietà causale falsificabile con dati pre-trattamento multipli (test dei pre-trend) ma non verificabile in modo definitivo. Il metodo del controllo sintetico di Abadie e Gardeazabal (2003) generalizza il DiD costruendo un gruppo di controllo come combinazione ponderata di unità di confronto, con pesi scelti per minimizzare la discrepanza pre-trattamento nelle variabili di esito — approssimando più finemente l'ipotetico controfattuale del gruppo trattato in assenza di intervento.

Tutti questi metodi condividono la struttura epistemica di un'approssimazione all'intervento puro: identificano l'effetto causale sotto assunzioni che restringono la famiglia dei modelli causali compatibili con i dati osservazionali, riducendo — ma non eliminando — l'ambiguità della classe di equivalenza di Markov. La loro differenza rispetto all'esperimento randomizzato è una differenza di grado, non di tipo: la randomizzazione elimina il confondimento per costruzione rendendo T indipendente da U per definizione, mentre i metodi quasi-sperimentali ottengono lo stesso risultato sotto assunzioni strutturali che devono essere motivate dalla teoria del dominio.

Il framework dei risultati potenziali di Rubin-Neyman: fondamenti matematici

Un elemento cruciale per la comprensione del problema causale, non sempre adeguatamente sviluppato nelle trattazioni che privilegiano il formalismo degli SCM di Pearl, e il confronto sistematico con la tradizione dei potential outcomes (risultati potenziali) di Neyman e Rubin. Questa tradizione --- matematicamente distinta nel formalismo ma equivalente nel contenuto al livello-3 della gerarchia causale --- è dominante in econometria, in epidemiologia e nell'analisi delle politiche pubbliche, e offre una prospettiva complementare che illumina aspetti del problema causale che il formalismo grafico non rende immediatamente visibili.

1.1 Assiomatica del framework PO

Definizione (Potential Outcome, Neyman 1923, Rubin 1974). Sia $U = \{1, \dots, N\}$ un insieme di unità sperimentali e sia T il insieme dei trattamenti, con $T = \{0, 1\}$ nel caso binario. Per ogni unità i in U e ogni trattamento t in T , si definisce $Y_i(t)$ come il risultato potenziale dell'unità i sotto il trattamento t --- il valore che la variabile di esito Y avrebbe assunto per l'unità i se essa avesse ricevuto il trattamento t . L'effetto causale individuale del trattamento per l'unità i è:

$$\tau_i = Y_i(1) - Y_i(0)$$

Il problema fondamentale dell'inferenza causale (Holland, 1986) è che per ogni unità i si osserva al più uno dei due risultati potenziali: se i riceve il trattamento 1, si osserva $Y_i(1)$ ma non $Y_i(0)$; se riceve il trattamento 0, si osserva $Y_i(0)$ ma non $Y_i(1)$. Il risultato non osservato è detto controfattuale. L'effetto causale individuale τ_i è quindi intrinsecamente non identificabile da una sola osservazione per unità.

SUTVA (Stable Unit Treatment Value Assumption, Rubin 1980). Il framework PO è fondato su due componenti della SUTVA:

(SUTVA-1) Assenza di interferenza: il risultato potenziale di ogni unità i non dipende dal trattamento assegnato ad altre unità $j \neq i$. Formalmente:

$$Y_i(t) = Y_i(t, T_{-i}) \text{ per ogni assegnazione } T_{-i} \text{ delle altre unit\textbf{a}}$$

dove T_{-i} indica il vettore dei trattamenti di tutte le unit\textbf{a} tranne i . Questa condizione esclude effetti di spillover, esternalità di rete, equilibri di mercato con prezzi endogeni, e dinamiche epidemiche.

(SUTVA-2) Unicit\textbf{a} della versione del trattamento: per ogni unit\textbf{a} i e ogni trattamento t , esiste un unico valore $Y_i(t)$, indipendentemente dal meccanismo o dal modo in cui il trattamento t viene somministrato. Questa condizione esclude l'ambiguit\textbf{a} sulla "dose" o sulla "forma" del trattamento.

La connessione con il formalismo degli SCM \textbf{e} diretta: SUTVA-1 corrisponde all'assunzione di modularit\textbf{a} degli SCM (ogni equazione strutturale \textbf{e} autonoma), \textbf{e} SUTVA-2 corrisponde all'interpretazione sharp del do-calculus ($\text{do}(X=x)$ specifica univocamente il valore di X). La violazione di SUTVA-1 richiede l'estensione al framework degli interference models, mentre la violazione di SUTVA-2 richiede il framework dei versioned treatments.

1.2 Gli estimandi causali e la loro struttura formale

Il framework PO distingue sistematicamente tra diversi estimandi causali --- quantit\textbf{a} che si intende stimare --- che corrispondono a domande causali diverse. Questa tipologia, pi\textbf{u} esplicita nel framework PO che negli SCM, \textbf{e} essenziale per comunicare con precisione in contesti applicativi.

Average Treatment Effect (ATE). L'effetto medio del trattamento nella popolazione \textbf{e}:

$$\text{ATE} = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

dove la seconda uguaglianza vale per la linearit\textbf{a} della speranza. Nel linguaggio degli SCM, $\text{ATE} = E[Y_{\{X=1\}}] - E[Y_{\{X=0\}}] = E[Y \mid \text{do}(X=1)] - E[Y \mid \text{do}(X=0)]$. L'ATE risponde alla domanda: qual \textbf{e} l'effetto medio del trattamento se lo si somministrasse uniformemente all'intera popolazione?

Average Treatment Effect on the Treated (ATT). L'effetto medio del trattamento sul gruppo che ha effettivamente ricevuto il trattamento \textbf{e}:

$$\text{ATT} = E[Y_i(1) - Y_i(0) \mid T_i = 1]$$

Nel linguaggio degli SCM, $\text{ATT} = E[Y_{\{X=1\}} - Y_{\{X=0\}} \mid X=1]$. L'ATT risponde alla domanda: qual \textbf{e} stato l'effetto del trattamento per coloro che lo hanno ricevuto? Questa \textbf{e} la quantit\textbf{a} rilevante per la valutazione delle politiche: se si vuole valutare l'impatto di un

programma di formazione professionale, si vuole sapere l'effetto sui partecipanti, non l'effetto medio ipoteticamente calcolato su tutti.

ATE e ATT differiscono quando la selezione al trattamento è correlata con l'eterogeneità del trattamento. Se individui con maggiore effetto del trattamento τ_i si autoselezionano al trattamento (positive selection on gains), allora $ATT > ATE$. In presenza di tale selezione, stimatori che producono stime dell'ATE (come i metodi di propensity score globali) sovrastimano o sottostimano l'ATT a seconda del segno della correlazione.

Local Average Treatment Effect (LATE, Imbens e Angrist, 1994). Il LATE, o Complier Average Causal Effect (CACE), è l'effetto causale del trattamento per la sotto-popolazione dei complier --- le unità per cui l'assegnazione strumentale $Z = 1$ (anziché $Z = 0$) causa effettivamente il trattamento $T_i = 1$ (anziché $T_i = 0$). Formalmente, i complier sono le unità i per cui:

$$Y_i(Z=1, \text{complier}) \text{ e tale che } T_i(Z=1) = 1 \text{ e } T_i(Z=0) = 0$$

Il LATE è definito come:

$$LATE = E[Y_i(T=1) - Y_i(T=0) \mid T_i(Z=1) = 1, T_i(Z=0) = 0]$$

Il teorema di Imbens e Angrist (1994) --- dimostrato sotto le assunzioni IV che saranno analizzate nella sezione sui metodi quasi-sperimentali --- stabilisce che lo stimatore IV identifica il LATE, non l'ATE, quando l'effetto del trattamento è eterogeneo tra le unità. Questa distinzione è cruciale: interpretare una stima IV come ATE è un errore concettuale quando vi è eterogeneità del trattamento.

Conditional Average Treatment Effect (CATE). Il CATE è l'effetto medio del trattamento condizionato su caratteristiche osservabili $Z = z$ delle unità:

$$CATE(z) = E[Y_i(1) - Y_i(0) \mid Z_i = z]$$

Nel linguaggio degli SCM, $CATE(z) = E[Y_{\{X=1\}} - Y_{\{X=0\}} \mid Z=z]$. Il CATE cattura l'eterogeneità dell'effetto del trattamento --- il fatto che lo stesso intervento può avere effetti diversi per sottogruppi diversi della popolazione. La stima del CATE è l'obiettivo del campo dell'individualized treatment rules e del precision medicine, ed è una delle aree di più intensa ricerca metodologica recente.

1.3 Il teorema di equivalenza PO-SCM

Teorema (equivalenza PO-SCM, Pearl 2000; Richardson e Robins 2013). Sia $M = \langle U, V, F, P(U) \rangle$ un SCM con grafo aciclico G , e sia $Y_x(u)$ il controfattuale di Y per l'unità u sotto l'intervento $do(X=x)$. Allora:

- (a) $Y_{x(u)} = Y_i(x)$ nel framework PO per l'unità i con caratteristica u , per ogni u e x .
- (b) $P(Y_x) = P(Y(x))$ per ogni distribuzione di U .
- (c) ATE nel framework PO = $E[Y \mid \text{do}(X=1)] - E[Y \mid \text{do}(X=0)]$ nel framework SCM.
- (d) ATT nel framework PO = $E[Y_{\{X=1\}} - Y_{\{X=0\}} \mid X=1]$ nel framework SCM.
- (e) CATE(z) nel framework PO = $E[Y_{\{X=1\}} - Y_{\{X=0\}} \mid Z=z]$ nel framework SCM.

La prova dell'equivalenza si basa sul fatto che l'intervento $\text{do}(X=x)$ nel framework SCM corrisponde esattamente all'assegnazione forzata $T_i = x$ nel framework PO: in entrambi i casi, si considera il meccanismo generatore dei dati con l'equazione strutturale di X rimossa e X fissato al valore x . I Single World Intervention Graphs (SWIG) di Richardson e Robins (2013) offrono una rappresentazione grafica che unifica i due framework, rendendo esplicita la corrispondenza tra i nodi Y_x del SWIG e i risultati potenziali $Y(x)$ del framework PO.

La differenza metodologica tra PO e SCM. I due framework non sono identici nella loro enfasi metodologica. Il framework PO privilegia l'esperimento randomizzato come via maestra all'inferenza causale: la randomizzazione del trattamento garantisce $T_i \perp (Y_i(0), Y_i(1))$, il che implica $E[Y \mid T=1] = E[Y(1)]$ e $E[Y \mid T=0] = E[Y(0)]$, rendendo l'ATE identificabile dalla semplice differenza delle medie osservate. Il framework SCM privilegia invece la modellazione strutturale del meccanismo generatore: il grafo rende esplicite le dipendenze strutturali, il do-calculus fornisce le condizioni di identificabilità per dati osservazionali, e la teoria delle classi di equivalenza quantifica l'ambiguità irriducibile.

I due framework si integrano reciprocamente: il PO è potente per l'analisi degli esperimenti (soprattutto in presenza di non-compliance e missing data), per la comunicazione delle assunzioni di identificabilità in termini di ignorabilità, e per la definizione precisa degli estimandi. Gli SCM sono potenti per la modellazione strutturale, la scoperta causale osservazionale, la dimostrazione dei teoremi di impossibilità, e per il ragionamento su sistemi complessi con molte variabili interagenti.

II. I metodi quasi-sperimentali come approssimazioni dell'intervento: analisi formale

La distinzione tra dati osservazionali e dati interventistici rischia di apparire come una dicotomia netta tra esperimento randomizzato (perfettamente interventistico) e osservazione passiva (puramente associativa). La pratica scientifica conosce invece una gamma continua di disegni di ricerca che si collocano tra questi estremi. I metodi quasi-sperimentali sfruttano variazioni naturali nel meccanismo di assegnazione al trattamento per approssimare la struttura interventistica. La loro analisi formale nel linguaggio degli SCM rivela sia la loro forza relativa rispetto alla pura osservazione sia i residui di ambiguità che li separano dall'esperimento ideale.

II.1 Il metodo delle variabili strumentali: fondamenti formali

Il metodo delle variabili strumentali (IV) è quello che ammette la formulazione più precisa nel linguaggio della teoria causale. Le sue origini risalgono a Wright (1928) nello studio dei mercati agricoli, e la sua sistematizzazione moderna si deve a Angrist, Imbens e Rubin (1996).

Definizione (strumento nel linguaggio degli SCM). Sia G il grafo causale con variabile di trattamento X , variabile di esito Y e confondimento latente U . Una variabile Z è uno strumento valido per la relazione causale $X \rightarrow Y$ se e solo se Z soddisfa le tre condizioni seguenti nel grafo causale:

(IV1) Rilevanza: $Z \rightarrow X$ in G (Z causa direttamente X).

(IV2) Esclusione: non esiste percorso diretto da Z a Y che aggiri X nel grafo.

(IV3) Esogeneità: $Z \perp\!\!\!\perp U$ in $P(U)$ (Z è indipendente dai confondenti latenti).

Nel grafo, queste condizioni richiedono che Z sia un genitore diretto di X , che non vi sia arco da Z a Y (né percorsi da Z a Y che bypassino X), e che non vi siano archi bidirezionali (indicatori di confondimento latente) tra Z e Y , né tra Z e U .

Identificabilità sotto linearità. Nel modello lineare con confondimento additivo:

$$X = \alpha * Z + \gamma * U + \epsilon_X$$

$$Y = \beta * X + \delta * U + \epsilon_Y$$

dove $\epsilon_X \perp\!\!\!\perp \epsilon_Y \perp\!\!\!\perp Z \perp\!\!\!\perp U$ e le variabili hanno media zero, lo stimatore IV identifica il coefficiente causale β come:

$$\beta_{IV} = \text{Cov}(Z, Y) / \text{Cov}(Z, X)$$

Derivazione. Per la condizione di rilevanza, $\text{Cov}(Z, X) = \alpha * \text{Var}(Z) \neq 0$. Per la condizione di esclusione ed esogeneità, $\text{Cov}(Z, Y) = \text{Cov}(Z, \beta * X + \delta * U + \epsilon_Y) = \beta * \text{Cov}(Z, X) + \delta * \text{Cov}(Z, U) + \text{Cov}(Z, \epsilon_Y)$. Per la condizione di esogeneità, $\text{Cov}(Z, U) = 0$, e per esogeneità di ϵ_Y da Z , $\text{Cov}(Z, \epsilon_Y) = 0$. Quindi $\text{Cov}(Z, Y) = \beta * \text{Cov}(Z, X)$, e $\beta_{IV} = \beta$. QED.

Identificabilità sotto eterogeneità (Imbens e Angrist, 1994). Nel modello non lineare con eterogeneità del trattamento (diverso effetto causale per unità diverse), lo stimatore IV identifica il LATE, non l'ATE. Questo risultato --- controintuitivo ma fondamentale --- segue dalla rappresentazione PO del modello IV. Con Z binario ($Z \in \{0, 1\}$) e T binario ($T \in \{0, 1\}$), si partizionano le unità in quattro tipi secondo il loro comportamento rispetto allo strumento:

- Complier: $T_i(Z=1) = 1$ e $T_i(Z=0) = 0$ (prende il trattamento se e solo se $Z=1$)
- Always-taker: $T_i(Z=1) = T_i(Z=0) = 1$ (prende sempre il trattamento)
- Never-taker: $T_i(Z=1) = T_i(Z=0) = 0$ (non prende mai il trattamento)
- Defier: $T_i(Z=1) = 0$ e $T_i(Z=0) = 1$ (prende il trattamento se e solo se $Z=0$)

Sotto l'assunzione di monotonicità (no defiers), il teorema di Imbens-Angrist stabilisce:

$$\text{LATE} = E[Y(1) - Y(0) \mid \text{complier}] = (E[Y \mid Z=1] - E[Y \mid Z=0]) / (E[T \mid Z=1] - E[T \mid Z=0])$$

Il denominatore è la frazione di complier nella popolazione, e il numeratore è la ridotta stimata dalla differenza nelle medie dell'esito tra i gruppi $Z=1$ e $Z=0$. Lo stimatore IV standard calcola esattamente il LATE sotto monotonicità e le tre condizioni IV.

Il limite non testabile: la condizione di esclusione. La condizione IV2 --- che Z non influenzi Y attraverso percorsi che bypassino X --- è l'assunzione critica non testabile della metodologia IV. Nel linguaggio degli SCM, è l'assunzione che Z non sia genitore diretto di Y . La sua violazione produce un bias nella stima del coefficiente causale:

$$\text{plim } \beta_{IV} = \beta + \text{Cov}(Z, Y_{\text{direct}}) / \text{Cov}(Z, X)$$

dove Y_{direct} indica l'effetto diretto di Z su Y che non passa per X . Uno strumento con violazione lieve della condizione di esclusione ($|\text{Cov}(Z, Y_{\text{direct}})|$ piccolo rispetto a $|\text{Cov}(Z, X)|$) produce un bias proporzionale alla debolezza dello strumento: per strumenti deboli ($\text{Cov}(Z, X)$ piccolo), anche una violazione minima dell'esclusione genera bias grandi. Questo è il problema degli strumenti deboli (weak instruments), formalizzato da Staiger e Stock (1997), che raccomandano come test pratico che la statistica F del primo stadio superi 10.

II.2 La regressione con discontinuità: analisi nel linguaggio causale

La discontinuità di regressione (RDD, Regression Discontinuity Design) sfrutta il fatto che l'assegnazione al trattamento dipende in modo discontinuo da una variabile di forza (running variable) attorno a una soglia critica. Il fondamento causale è un'assunzione locale di ignorabilità attorno alla soglia.

Disegno formale. Sia R la running variable (variabile di forza), c la soglia critica, e $T = 1[R \geq c]$ l'assegnazione al trattamento. L'RDD sharp assume che non vi sia manipolazione strategica di R attorno a c : per unità con R vicino a c , l'assegnazione al trattamento è quasi-casuale, perché la deviazione da c può essere attribuita a fattori esogeni.

Nel linguaggio degli SCM. La condizione di identificabilità dell'RDD è che il meccanismo che genera R sia esogeno rispetto ai confondenti U nel vicinato della soglia --- ossia che la distribuzione dei confondenti sia continua attorno a c :

$$E[U \mid R = c+] = E[U \mid R = c-] \quad (\text{continuità dei confondenti alla soglia})$$

Sotto questa condizione, la differenza nelle medie dell'esito ai due lati della soglia identifica il LATE al valore soglia:

$$LATE(c) = E[Y(1) - Y(0) \mid R = c] = \lim_{r \rightarrow c+} E[Y \mid R=r] - \lim_{r \rightarrow c-} E[Y \mid R=r]$$

L'assunzione di identificabilità è locale: si stima l'effetto causale per le unità con R vicino a c , non per l'intera popolazione. La generalizzabilità del $LATE(c)$ all'ATE richiede assunzioni aggiuntive sull'omogeneità dell'effetto del trattamento rispetto al valore di R .

Test di validità. L'assunzione di continuità dei confondenti alla soglia non è direttamente testabile, ma può essere parzialmente saggiata con:

(a) McCrary density test: se le unità manipolano R per posizionarsi appena sopra la soglia (anticipando i benefici del trattamento), la densità di R è discontinua in c . Un test di densità (McCrary, 2008) testa la discontinuità della densità campionaria di R in c .

(b) Continuità delle covariate predeterminate: le variabili predeterminate (caratteristiche baseline che non possono essere causate dal trattamento) dovrebbero avere distribuzioni continue in c . La discontinuità di una covariata predeterminata in c è evidenza contro la validità dell'RDD.

II.3 Differenze-nelle-differenze: struttura causale e ipotesi di identificabilità

Il metodo delle differenze-nelle-differenze (DiD) sfrutta la variazione temporale del trattamento: un gruppo trattato passa da $T=0$ a $T=1$ in un momento t^* , mentre un gruppo di controllo rimane sempre a $T=0$. L'assunzione causale fondamentale è l'assunzione delle tendenze parallele.

Assunzione delle tendenze parallele (nel linguaggio causale). In assenza del trattamento, il gruppo trattato avrebbe seguito la stessa tendenza temporale del gruppo di controllo. Nel linguaggio PO:

$$E[Y_i(0, t) - Y_i(0, t-1) \mid D_i = 1] = E[Y_i(0, t) - Y_i(0, t-1) \mid D_i = 0]$$

dove $D_i = 1$ indica l'appartenenza al gruppo trattato e $Y_i(0, t)$ è il risultato potenziale dell'unità i al tempo t senza trattamento. Nel linguaggio degli SCM, questa ipotesi equivale a richiedere che la distribuzione del confounder latente U , condizionata al gruppo, sia stazionaria nel tempo --- una condizione di stazionarietà causale.

Identificabilità dello stimatore DiD. Sotto l'assunzione di tendenze parallele, lo stimatore DiD identifica l'ATT:

$$ATT = (E[Y \mid D=1, t=t^*] - E[Y \mid D=1, t=t^*-1]) - (E[Y \mid D=0, t=t^*] - E[Y \mid D=0, t=t^*-1])$$

La derivazione segue direttamente: la prima differenza ($E[Y_{\{D=1, t^*\}}] - E[Y_{\{D=1, t^*-1\}}]$) cattura sia l'effetto del trattamento sia la tendenza temporale per il gruppo trattato; la seconda differenza ($E[Y_{\{D=0, t^*\}}] - E[Y_{\{D=0, t^*-1\}}]$) cattura solo la tendenza temporale per il gruppo di controllo. La loro differenza, sotto tendenze parallele, isola l'effetto causale del trattamento.

Il test dei pre-trend. L'assunzione di tendenze parallele è parzialmente testabile con dati pre-trattamento multipli: se prima del trattamento i due gruppi mostrano tendenze diverse, l'assunzione è violata anche nel periodo post-trattamento. Il test dei pre-trend (test di uguaglianza dei trend nei periodi pre-trattamento) è una condizione necessaria ma non sufficiente per la validità del DiD: la parallelità pre-trattamento non garantisce la parallelità controfattuale post-trattamento.

DiD con staggered treatment adoption. Nella pratica econometrica moderna, il DiD è spesso applicato a contesti in cui il trattamento viene adottato in tempi diversi da unità diverse (staggered adoption). Callaway e Sant'Anna (2021) e Sun e Abraham (2021) hanno

dimostrato che lo stimatore DiD con effetti fissi bidirezionali (two-way fixed effects) non identifica una media ponderata degli ATT in presenza di eterogeneità del trattamento con staggered adoption, ma può produrre stime con segno opposto all'effetto causale vero. I nuovi estimatori DiD con staggered adoption correggono questo problema decomponendo l'estimando per coorte e periodo di trattamento.

II.4 Il controllo sintetico: costruzione del controfattuale

Il metodo del controllo sintetico (Abadie e Gardeazabal, 2003; Abadie, Diamond e Hainmueller, 2010) affronta il problema del DiD nei casi in cui il gruppo di controllo naturale non è comparabile al gruppo trattato nei periodi pre-trattamento. La soluzione è costruire un gruppo di controllo sintetico come combinazione convessa ponderata delle unità di controllo disponibili, con pesi scelti per minimizzare la discrepanza pre-trattamento nelle variabili di esito e nelle caratteristiche predeterminate.

Formalizzazione. Sia $i=1$ l'unità trattata (tratta il trattamento in $t=T^*$) e siano $i=2, \dots, J+1$ le J unità di controllo. I pesi del controllo sintetico $W^* = (w_2, \dots, w_{J+1})$ con $w_j \geq 0$ e $\sum w_j = 1$ sono definiti come:

$$W^* = \operatorname{argmin}_W \sum_{\{t < T^*\}} (Y_{1t} - \sum_{j=2}^J w_j Y_{jt})^2 + \lambda \sum_{j=2}^J |X_{1t} - \sum_{j=2}^J w_j X_{jt}|^2$$

dove Y_{it} è la variabile di esito e X_i è il vettore di caratteristiche predeterminate. Il controfattuale stimato per l'unità trattata in assenza di trattamento è:

$$Y_1(0, t) = \sum_{j=2}^J w_j Y_{jt} \quad \text{per } t \geq T^*$$

e l'effetto causale stimato per ogni periodo post-trattamento $t \geq T^*$ è:

$$\tau(t) = Y_{1t} - \sum_{j=2}^J w_j Y_{jt}$$

Condizioni di identificabilità nel linguaggio degli SCM. L'assunzione di identificabilità del controllo sintetico è che il controfattuale $Y_1(0, t)$ per $t \geq T^*$ possa essere approssimato da una combinazione lineare delle unità di controllo. Questa assunzione --- più debole dell'assunzione di tendenze parallele del DiD, che richiede che la serie completa del gruppo trattato sia approssimabile da una sola unità di controllo --- è l'assunzione di identificabilità chiave. Nel linguaggio degli SCM, equivale a richiedere che il confondente latente U del gruppo trattato sia generato dallo stesso fattore comune che guida le unità di controllo, in modo che la combinazione ponderata delle unità di controllo approssimi il controfattuale.

Inferenza: permutation tests. L'inferenza statistica nel controllo sintetico non può usare i test standard (la varianza del CPSC è zero per costruzione nel pre-periodo), e si basa su permutation tests: si applica la stessa procedura di stima a ogni unità di controllo come se fosse trattata nello stesso momento, ottenendo una distribuzione delle discrepanze pre-post per le unità non trattate. L'effetto stimato per l'unità trattata è significativo se è un outlier rispetto a questa distribuzione di placebo.

III. La struttura epistemica dei metodi quasi-sperimentali nel contesto del framework causale

Tutti i metodi quasi-sperimentali condividono la struttura epistemica di un'approssimazione all'intervento puro, e la loro analisi nel linguaggio degli SCM chiarisce sia la loro forza sia i loro limiti fondamentali.

Gerarchia epistemica. I metodi quasi-sperimentali possono essere ordinati per la fedeltà con cui approssimano un intervento causale puro:

1. Esperimento randomizzato: eliminazione del confondimento per costruzione. $T_{-} | T_{+} \perp U$ per definizione. Identificabilità dell'ATE senza assunzioni.
2. IV: eliminazione del confondimento attraverso la variazione esogena dello strumento. Identificabilità del LATE sotto assunzioni non testabili (esclusione, esogeneità, monotonicità).
3. RDD: eliminazione locale del confondimento attraverso la continuità dei confondenti alla soglia. Identificabilità del LATE(c) sotto assunzioni locali non testabili.
4. DiD: eliminazione del confondimento invariante nel tempo attraverso la differenziazione temporale. Identificabilità dell'ATT sotto assunzioni di tendenze parallele parzialmente testabili.
5. Controllo sintetico: costruzione del controfattuale per una singola unità. Identificabilità dell'ATT sotto assunzioni di fattore comune.
6. Regressione con controlli: riduzione del confondimento attraverso il condizionamento su covariate osservate. Identificabilità dell'ATE sotto assunzione di ignorabilità condizionale non testabile.
7. Dati osservazionali non aggiustati: nessuna eliminazione del confondimento.

Il confine comune. Tutti questi metodi condividono un confine fondamentale rispetto all'esperimento randomizzato: le assunzioni di identificabilità non sono verificabili dai dati stessi, ma devono essere motivate da conoscenza teorica del dominio. Nel linguaggio degli SCM, ogni metodo quasi-sperimentale assume implicitamente un grafo causale con determinate proprietà (assenza di certi archi, indipendenza tra certi nodi), e la sua validità dipende dalla correttezza di quel grafo. La differenza rispetto all'osservazione passiva e quantitativa --- i metodi quasi-sperimentali restringono la famiglia dei grafi compatibili con i dati --- non qualitativa: non possono mai eliminare completamente l'ambiguità causale senza informazione esogena al di là dei dati.

Implicazione per l'intelligenza artificiale. L'analisi dei metodi quasi-sperimentali nel contesto del framework causale chiarisce perché i sistemi di apprendimento automatico addestrati su dati osservazionali non possono apprendere relazioni causali nel senso di Pearl o di Rubin, indipendentemente dalla loro architettura o dalla dimensione del corpus di addestramento. L'addestramento su dati osservazionali -- anche di enorme scala -- corrisponde al caso 7 della gerarchia epistemica sopra, il più lontano dall'intervento causale. I metodi quasi-sperimentali mostrano che anche avvicinarsi all'intervento causale richiede assunzioni sostanziali sul meccanismo generatore dei dati che non possono essere estratte dai dati stessi. Questa è la conclusione epistemica fondamentale del framework causale applicata ai sistemi di IA.

La struttura del confine e il ciclo di monografie

Il ciclo di monografie al quale questa si aggiunge ha esplorato sistematicamente i limiti di ciò che può essere conosciuto dall'esterno di un sistema a partire dalle sue manifestazioni. I risultati di Gödel e Turing hanno mostrato che nessun sistema formale sufficientemente potente può decidere dall'interno la propria consistenza e che il problema della terminazione di un programma non è decidibile da nessun algoritmo universale. La geometria di Fisher-Rao e la teoria dell'informazione di Shannon hanno precisato quanto una compressione può catturare di una sorgente: l'entropia è il confine inferiore della compressione lossless, e la capacità del canale è il confine superiore della comunicazione affidabile. L'impossibilità di Arrow ha delimitato la distanza tra preferenze individuali e scelta collettiva razionale; il degrado di Goodhart ha mostrato come l'ottimizzazione di una misura distrugga la misura come indicatore del costrutto che si voleva misurare. Il problema della coscienza e il test di Turing hanno tracciato il confine tra vita interna e comportamento osservabile; la teoria dei giochi e il meccanismo design hanno circoscritto le possibilità dell'intelligenza collettiva; il dibattito tra connessionismo e logica ha ridefinito il ruolo della certificazione formale nell'era delle reti neurali profonde.

Questa settima monografia aggiunge un tassello che le precedenti non hanno esaminato in modo diretto, pur avendone preparato il terreno concettuale: il confine tra il livello dell'associazione e il livello dell'intervento nella gerarchia della conoscenza causale. Il risultato fondamentale può essere enunciato in forma preliminare con queste parole: nessuna quantità di dati osservazionali, per quanto sterminata e accurata, può per sé sola fondare un'inferenza sulle distribuzioni che si otterrebbero intervenendo attivamente sul sistema generatore di

quei dati, salvo che si assumano ipotesi sulla struttura causale di quel sistema — ipotesi che i dati stessi non possono convalidare senza incorrere in una circolarità. Questo non è un risultato intuitivo che attendeva una dimostrazione formale: è un teorema, con premesse precise e conseguenze che si dispiegano attraverso la teoria dell'identificabilità, la teoria della scoperta causale e l'analisi delle architetture di apprendimento automatico che costituiscono oggi il nucleo operativo dell'intelligenza artificiale.

La relazione di questo confine con quelli esplorati nei volumi precedenti è istruttiva. Il confine gödeliano separa la verità matematica dalla derivabilità in un sistema formale dato: vi sono verità aritmetiche che non possono essere dimostrate in nessun sistema assiomatico completo e consistente sufficientemente potente. Il confine di Turing separa i problemi decidibili da quelli indecidibili: vi sono domande matematicamente precise a cui nessun algoritmo può rispondere per tutti gli input in tempo finito. Il confine causale separa la conoscenza osservazionale dalla conoscenza interventistica: vi sono fatti causali sul mondo che nessuna accumulazione di dati passivi può determinare, per quante osservazioni si raccolgano. La struttura dei tre confini è analoga: in tutti e tre i casi si ha una gerarchia stretta, con livelli non collassabili, separati da una distanza che è logica prima ancora che pratica.

Tre obiettivi e struttura dell'argomentazione

La monografia si propone tre obiettivi distinti ma convergenti.

Il primo obiettivo è ricostruire la struttura formale del confine nella sua massima generalità: enunciare con precisione matematica in quale senso la gerarchia causale è *stretta* — cioè in quale senso nessuna funzione calcolabile può attraversare i livelli della gerarchia partendo dal solo livello osservazionale. Questo richiede di formalizzare la nozione di modello causale strutturale, di definire l'operatore di intervento come operazione di mutilazione del grafo, e di dimostrare che la classe dei modelli compatibili con una distribuzione osservazionale è sempre non degenerare — contiene sempre più di un modello quando vi è confondimento latente, e i modelli in quella classe possono produrre distribuzioni post-intervento arbitrariamente diverse.

Il secondo obiettivo è esplorare le condizioni sotto cui quel confine può essere parzialmente valicato: la teoria dell'identificabilità causale costruisce le condizioni necessarie e sufficienti perché una quantità interventistica sia calcolabile dalla distribuzione osservazionale, dato che il grafo causale qualitativo è noto. Questo include il criterio del backdoor, il criterio del front-door, il do-calculus di Pearl nella sua formulazione in tre regole e il teorema di completezza di Shpitser e Pearl, che trasforma il sistema da euristico ad algoritmo di decisione. Si esplorano poi i limiti intrinseci di questa identificabilità parziale: i grafi per cui il do-calculus si arresta, le c-componenti, il teorema di Tian e Pearl sulla caratterizzazione geometrica della non-identificabilità attraverso la nozione di hedge.

Il terzo obiettivo — il più direttamente rilevante per il dibattito contemporaneo sull'intelligenza artificiale — è trarre le conseguenze architettureali di questi teoremi per i sistemi di apprendimento automatico. Il paradigma dell'addestramento per minimizzazione della perdita su distribuzioni fisse genera modelli che approssimano funzioni della distribuzione osservazionale: macchine di livello-1 nella gerarchia di Pearl. Questo confinamento non è un difetto contingente correggibile con più dati o architetture più ampie,

ma una conseguenza necessaria del modo in cui vengono addestrati. Le capacità apparentemente causali che questi sistemi esibiscono — la capacità di rispondere correttamente a domande causali, di completare ragionamenti controfattuali, di identificare meccanismi — sono il prodotto di correlazioni statistiche di ordine superiore molto dense e strutturate, non di modelli causali interni genuini. La distinzione ha conseguenze pratiche precise: riguarda la robustezza dei sistemi sotto shift distributivi causati da interventi, la correttezza delle raccomandazioni in contesti decisionali e la validità delle valutazioni comparative tra sistemi.

Note metodologiche preliminari

Prima di procedere è opportuno chiarire alcune convenzioni metodologiche che strutturano la trattazione.

La distinzione tra limiti tecnici e limiti di principio attraversa l'intera monografia come linea di demarcazione fondamentale. I limiti *computazionali* — come la NP-difficoltà dell'apprendimento strutturale esatto o la dipendenza esponenziale dalla dimensione del set di condizionamento nei test di indipendenza non parametrici — sono problemi che la crescita delle risorse computazionali può alleviare ma non risolvere, e che algoritmi più ingegnosi possono aggirare parzialmente. I limiti *di principio* — come la non-identificabilità delle distribuzioni interventistiche in presenza di confondimento latente non controllato, o la non-identificabilità dei controfattuali anche quando tutte le distribuzioni interventistiche sono note — sono impossibilità logiche nel senso precisato sopra: non migliorabili con nessuna quantità di risorse o sofisticazione algoritmica. La confusione tra questi due livelli è uno degli errori più frequenti nel dibattito sull'IA causale, e questa monografia si propone di dissiparla sistematicamente.

Riguardo al rapporto con i volumi precedenti di questo ciclo: i modelli causali strutturali di Pearl, l'operatore *do*, il *do*-calculus nella sua formulazione in tre regole, l'entropia di trasferimento e la sua derivazione come informazione mutua condizionale sono stati costruiti con rigore nel quinto volume. Questa monografia li presuppone come strumenti noti, li richiama quando necessario e rimanda alle derivazioni già condotte, senza ripeterle per esteso. L'obiettivo qui è usare il formalismo come punto di partenza per esplorare i confini dell'identificabilità e le conseguenze architetturali per i sistemi artificiali — un territorio che il volume precedente aveva lasciato deliberatamente intatto.

La struttura in sei parti

La Prima Parte costruisce il problema a partire dal paradosso di Simpson, mostrandone la natura non come anomalia statistica bensì come teorema di struttura: la coesistenza di $P(Y|X=1) > P(Y|X=0)$ con $P(Y|X=1,Z=z) < P(Y|X=0,Z=z)$ per ogni z rivela che le relazioni di condizionamento statistico non si lasciano interpretare causalmente senza assumere struttura causale aggiuntiva — struttura che i dati stessi non contengono. La fenomenologia del paradosso — le sue tre risoluzioni causali incompatibili, ognuna ugualmente compatibile con i dati — è il caso paradigmatico più semplice della non-identificabilità causale. La Prima Parte conclude con la formulazione del teorema di incommensurabilità tra livello-1 e livello-2 della gerarchia, che anticipa i risultati più generali delle Parti successive.

La Seconda Parte formalizza la gerarchia causale come sistema di teoremi di impossibilità e ne deriva la strettezza in senso tecnico preciso. La dimostrazione procede per stadi: prima il teorema fondamentale di non-identificabilità in presenza di confondimento latente, con costruzione esplicita della famiglia di modelli compatibili; poi la teoria dell'identificabilità condizionale al grafo, con la formula di aggiustamento per backdoor e la sua derivazione rigorosa dal formalismo della mutilazione; poi il do-calculus e il suo teorema di completezza; infine la caratterizzazione della non-identificabilità attraverso c-componenti e hedge. La transizione al livello-3 — il confine tra distribuzione interventistica e controfattuali — viene trattata separatamente, poiché richiede strumenti diversi e la sua struttura è fondamentalmente distinta.

La Terza Parte sviluppa la teoria dell'equivalenza di Markov e la topologia dello spazio dei modelli causali compatibili con una distribuzione osservazionale. Il teorema di Verma caratterizza le classi di equivalenza attraverso scheletro e v-strutture; il CPDAG fornisce la rappresentazione canonica unica di ogni classe; le regole di Meek forniscono l'algoritmo di costruzione. La geometria dello spazio dei CPDAG — le dimensioni delle classi di equivalenza, la riduzione dell'ambiguità con dati interventistici secondo i teoremi di Eberhardt — quantifica l'ambiguità causale irriducibile nei dati osservazionali come problema di struttura, non di informazione incompleta. La Parte include inoltre la trattazione dei Maximal Ancestral Graphs e dei Partial Ancestral Graphs, che estendono il formalismo al caso con variabili latenti arbitrarie.

La Quarta Parte esamina gli algoritmi di scoperta causale con le condizioni matematiche precise sotto cui estraggono informazione causale da dati osservazionali. L'algoritmo PC e la sua versione PC-stable ad alta dimensione, l'algoritmo GES basato su punteggi di verosimiglianza, LINGAM con il fondamento nell'analisi delle componenti indipendenti, i modelli a rumore additivo bivariati. Una trattazione sistematica è dedicata ai test di indipendenza condizionale kernel-based (HSIC, KCI), che consentono di operare in spazi non parametrici mantenendo garanzie statistiche formali. Per ciascun algoritmo si esplicitano le assunzioni, le garanzie di consistenza, le stime della dimensione campionaria necessaria e i limiti non superabili. Il teorema di Chickering sulla NP-difficoltà chiude la parte sul versante computazionale.

La Quinta Parte sviluppa la diagnosi strutturale dei sistemi di IA come macchine associative. Il paradigma dell'addestramento per minimizzazione della perdita su distribuzioni fisse viene formalizzato e il suo confinamento al livello-1 viene dimostrato come conseguenza della struttura dell'obiettivo di addestramento. I meccanismi dell'illusione di comprensione causale, il problema della generalizzazione fuori distribuzione e la sua connessione formale con la causalità, la riformulazione della legge di Goodhart come teorema causale, i tre meccanismi classici di distorsione causale nei sistemi di IA — correlazione spuria, causalità inversa, confondimento — vengono analizzati con i formalismi della teoria causale. L'analisi dell'architettura Transformer come macchina associativa formale chiude la sezione con una valutazione delle implicazioni per il disegno e la valutazione dei sistemi.

La Sesta Parte esamina i tentativi recenti di costruire rappresentazioni causali nei sistemi di apprendimento automatico, analizzandone le garanzie formali e i limiti strutturali residui. Il teorema di Locatello sull'impossibilità del disentanglement non supervisionato, i risultati di Khemakhem e Hyvärinen sull'identificabilità non lineare con supervisione debole, il teorema di Peters e Bühlmann sull'identificabilità dei modelli causali non lineari, il framework della

trasportabilità di Pearl, il contributo di Schölkopf e collaboratori sul Causal Representation Learning multi-ambiente. La Parte si conclude con l'enunciazione del limite fondamentale che nessun approccio all'interno del paradigma dell'apprendimento può superare senza dati interventistici, e con una mappa precisa delle possibilità in funzione del tipo di dati disponibili.

Le Appendici sviluppano in forma completa e autocontenuta le derivazioni che avrebbero appesantito il testo principale: le prove di soundness delle tre regole del do-calculus; le proprietà fondamentali della divergenza di Kullback-Leibler e la sua relazione con l'entropia di trasferimento; la riduzione da 3-colorazione di grafi per la NP-difficoltà dell'apprendimento strutturale; i principi di identificabilità nei modelli lineari gaussiani con le condizioni di rango econometriche classiche; la formula di trasporto e il teorema di completezza della trasportabilità; la teoria degli strumenti deboli e le correzioni di Bekker.

Parte Prima: Il Paradosso di Simpson e la Struttura Formale del Confine

I. Un'anomalia che non è un'anomalia

Il paradosso di Simpson — denominato in onore di Edward H. Simpson che ne offrì nel 1951 una trattazione sistematica sul *Journal of the Royal Statistical Society*, benché Karl Pearson nel 1899 e Udny Yule nel 1903 ne avessero già segnalato la struttura in contesti applicativi distinti — è uno dei risultati più citati e meno compresi dell'intera tradizione statistica. Viene presentato, nei manuali introduttivi, come un fenomeno curioso che ammaestra alla prudenza nell'interpretazione dei dati aggregati; viene evocato, nel giornalismo scientifico, come monito contro le conclusioni affrettate; viene talvolta impiegato, nella letteratura epistemologica sulla causalità, come motivazione per introdurre modelli causali espliciti. Tutte queste presentazioni colgono qualcosa di reale ma restano incomplete nella portata: il paradosso di Simpson non è un'anomalia che la cautela metodologica può aggirare, né un effetto statistico tra i molti che richiedono avvertenza. È un teorema di struttura — un risultato che afferma qualcosa di preciso sulla topologia delle relazioni di condizionamento e sulla loro relazione irriducibilmente ambigua con la struttura causale sottostante.

Karl Pearson nel 1899, nel suo articolo *Genetic (Reproductive) Selection*, osservò che la correlazione tra due variabili misurata su una popolazione aggregata può essere di segno opposto alla correlazione misurata all'interno di ciascun sottogruppo omogeneo della stessa popolazione. Egli interpretò il fenomeno come un problema di eterogeneità della popolazione senza tuttavia trarne conseguenze sistematiche per l'interpretazione causale. Yule, nel 1903, osservò lo stesso fenomeno in dati demografici sul pauperismo, notando che l'associazione tra la politica del lavoro forzato e il tasso di pauperismo si invertiva quando si stratificava per tipo di parrocchia, e riconobbe esplicitamente che il fenomeno poneva un problema interpretativo serio senza identificarne la soluzione strutturale.

Il contributo di Simpson del 1951 è distinto da questi precursori in un punto essenziale: egli fornisce la prima dimostrazione algebrica sistematica che l'inversione non è un artefatto di calcoli scorretti ma una proprietà matematicamente possibile delle distribuzioni di probabilità

condizionale, costruendo esempi numerici espliciti in cui le probabilità condizionali stratificate indicano uniformemente una direzione mentre la probabilità condizionale aggregata ne indica un'altra. La reinterpretazione causale del paradosso è tuttavia opera di Judea Pearl, che nella sua analisi del 1993 e poi nel volume *Causality* del 2000 mostrò che le tre risoluzioni del paradosso — stratificare, non stratificare, ignorare entrambe le misure — corrispondono a tre strutture causali distinte, ognuna compatibile con la stessa distribuzione osservazionale, e che la scelta tra esse non è una questione di preferenza metodologica ma di fatto causale sul meccanismo generatore dei dati.

II. Il teorema di Simpson nella sua forma più generale

Prima di illustrarne il contenuto con esempi, è opportuno enunciare il teorema nella sua forma algebrica più generale, poiché questa generalità mostra che il paradosso non è un fenomeno marginale che richiede condizioni molto specifiche per manifestarsi: è una proprietà strutturale delle relazioni di condizionamento che si realizza per una classe ampia e tipica di distribuzioni.

Teorema (Simpson, forma generale). Siano X, Y, Z variabili aleatorie a valori in spazi finiti e sia P una distribuzione di probabilità congiunta su (X, Y, Z) . Non è necessario in generale che la direzione dell'associazione marginale $P(Y|X=1)$ rispetto a $P(Y|X=0)$ concordi con la direzione dell'associazione condizionale $P(Y|X=1, Z=z)$ rispetto a $P(Y|X=0, Z=z)$ per ogni z . Più precisamente, è algebricamente realizzabile la seguente configurazione simultanea:

$$P(Y=1 | X=1) > P(Y=1 | X=0)$$

e contemporaneamente, per ogni valore z appartenente al supporto di Z :

$$P(Y=1 | X=1, Z=z) < P(Y=1 | X=0, Z=z)$$

La costruzione che segue dimostra il risultato per via esplicita con variabili binarie, scegliendo i valori in modo da rendere massimamente trasparente il meccanismo dell'inversione. Si consideri la distribuzione congiunta definita dai seguenti valori condizionali:

$$P(Z=1 | X=1) = 3/4, P(Z=0 | X=1) = 1/4$$

$$P(Z=1 | X=0) = 1/4, P(Z=0 | X=0) = 3/4$$

$$P(Y=1 | X=1, Z=1) = 0.70, P(Y=1 | X=0, Z=1) = 0.80$$

$$P(Y=1 | X=1, Z=0) = 0.20, P(Y=1 | X=0, Z=0) = 0.30$$

Applicando la formula della probabilità totale, che marginalizza su Z pesando le probabilità condizionali con la distribuzione di Z nel gruppo corrispondente:

$$P(Y=1 | X=1) = 0.70 \cdot (3/4) + 0.20 \cdot (1/4) = 0.525 + 0.050 = 0.575$$

$$P(Y=1 | X=0) = 0.80 \cdot (1/4) + 0.30 \cdot (3/4) = 0.200 + 0.225 = 0.425$$

Dunque $P(Y=1|X=1) = 0.575 > 0.425 = P(Y=1|X=0)$: la direzione marginale indica che $X=1$ è associato a probabilità maggiore di $Y=1$. Eppure per entrambi i valori di Z la relazione è uniformemente invertita: $0.70 < 0.80$ (per $Z=1$) e $0.20 < 0.30$ (per $Z=0$).

Il meccanismo è trasparente: il gruppo $X=1$ è concentrato per tre quarti nel sottogruppo $Z=1$, che è il sottogruppo con probabilità di base di $Y=1$ più elevata per entrambi i valori di X . Il gruppo $X=0$ è concentrato per tre quarti nel sottogruppo $Z=0$, con probabilità di base più bassa. La differenza di concentrazione — la correlazione positiva tra X e Z — domina le differenze stratificate tra X e Y all'interno di ciascuno strato, capovolgendo la direzione aggregata.

III. La struttura algebrica dell'inversione: decomposizione formale completa

Comprendere perché questa inversione si produce — e quando si verifica — richiede di esaminare l'operazione di marginalizzazione nella sua forma algebrica. La probabilità condizionale marginale è una media pesata delle probabilità condizionali stratificate, con pesi che dipendono dalla distribuzione di Z all'interno di ciascun gruppo. Introducendo la notazione abbreviata $\mu(x, z) = P(Y=1 | X=x, Z=z)$ e $w_x(z) = P(Z=z | X=x)$, si ha:

$$P(Y=1 | X=x) = \sum_z \mu(x, z) \cdot w_x(z)$$

La differenza marginale tra i due gruppi è:

$$\Delta_{\text{marg}} = P(Y=1 | X=1) - P(Y=1 | X=0) = \sum_z \mu(1, z) \cdot w_1(z) - \sum_z \mu(0, z) \cdot w_0(z)$$

Si aggiunge e si sottrae il termine di confronto $\sum_z \mu(0, z) \cdot w_1(z)$, ottenendo la decomposizione fondamentale:

$$\Delta_{\text{marg}} = \sum_z [\mu(1, z) - \mu(0, z)] \cdot w_1(z) + \sum_z \mu(0, z) \cdot [w_1(z) - w_0(z)]$$

Il primo termine — chiamato componente causale pesata (CCP) — è la media ponderata con i pesi del gruppo $X=1$ delle differenze stratificate $\delta(z) = \mu(1, z) - \mu(0, z)$. Se le differenze stratificate sono tutte negative ($\delta(z) < 0$ per ogni z), questo termine è negativo. Il secondo termine — chiamato componente di confondimento distributivo (CCD) — cattura la distorsione introdotta dalla differenza nelle distribuzioni di Z tra i due gruppi: è proporzionale a quanto la probabilità di base di Y varia tra strati e a quanto diversamente i due gruppi sono concentrati negli strati.

L'inversione di Simpson — il caso in cui $\Delta_{\text{marg}} > 0$ e $\delta(z) < 0$ per ogni z — si verifica quando il termine CCD è positivo e domina il termine CCP in valore assoluto:

$$\sum_z \mu(0, z) \cdot [w_1(z) - w_0(z)] > |\sum_z \delta(z) \cdot w_1(z)|$$

Una formulazione alternativa impiega la covarianza condizionale. Si può mostrare che:

$$\Delta_{\text{marg}} = \mathbb{E}[\delta(Z) | X=1] + \text{Cov}_{\{P(Z)\}}(\mu(0, Z), \Delta w(Z))$$

dove $\Delta w(z) = w_1(z) - w_0(z)$ è la differenza nella distribuzione di Z tra i due gruppi, e la covarianza è calcolata rispetto alla distribuzione marginale di Z . Il termine $\text{Cov}(\mu(0, Z), \Delta w(Z))$ è il prodotto scalare tra la funzione di probabilità di base $\mu(0, \cdot)$ e la funzione di squilibrio distributivo $\Delta w(\cdot)$. Questo termine è positivo quando gli strati con alta probabilità di base di Y coincidono con gli strati in cui il gruppo $X=1$ è sovrarappresentato rispetto al gruppo $X=0$ — esattamente la condizione dell'esempio numerico.

L'inversione è quindi facilitata da tre condizioni strutturali che si moltiplicano: una differenza forte nella distribuzione di Z tra i due gruppi (il termine $|\Delta w|$ grande), una correlazione elevata tra lo strato Z e la probabilità di base di Y nel gruppo di riferimento (la covarianza grande) e un effetto causale di X su Y relativamente debole in ciascuno strato (il termine $\mathbb{E}[\delta(Z)|X=1]$ piccolo in valore assoluto). Quando queste condizioni sono soddisfatte simultaneamente, il confondimento distributivo prevale e la direzione marginale diventa sistematicamente fuorviante rispetto all'effetto causale reale.

IV. Le tre risoluzioni causali e la loro struttura formale

Il paradosso ammette tre risoluzioni causali distinte, a seconda del ruolo che Z svolge nel meccanismo generatore dei dati. Insistiamo su un punto fondamentale: queste non sono opzioni metodologiche tra cui scegliere in base alla comodità o alla tradizione della disciplina. Sono affermazioni incompatibili su fatti causali, la cui verità dipende dalla struttura del meccanismo generatore dei dati e non dai dati medesimi. Qualunque statistica calcolabile da $P(X, Y, Z)$ è compatibile con ognuna delle tre strutture causali descritte di seguito.

Prima risoluzione: Z è una variabile di confondimento. Il meccanismo causale ha la struttura $X \leftarrow Z \rightarrow Y$, con un percorso causale diretto $X \rightarrow Y$. Intervenire su X modifica Y attraverso il percorso causale, ma la correlazione osservazionale include anche il contributo del percorso backdoor $X \leftarrow Z \rightarrow Y$. La quantità causalmente rilevante è la formula di standardizzazione:

$$P(Y=1 \mid \text{do}(X=1)) = \sum_z P(Y=1 \mid X=1, Z=z) \cdot P(Z=z)$$

Con distribuzione marginale $P(Z=1) = P(Z=0) = 1/2$ nell'esempio numerico precedente, si ottiene:

$$P(Y=1 \mid \text{do}(X=1)) = 0.70 \cdot (1/2) + 0.20 \cdot (1/2) = 0.45$$

$$P(Y=1 \mid \text{do}(X=0)) = 0.80 \cdot (1/2) + 0.30 \cdot (1/2) = 0.55$$

L'effetto causale di X su Y è negativo: intervenire per portare X da 0 a 1 riduce la probabilità di Y di 0.10. La differenza marginale positiva era interamente un artefatto di confondimento. Questo scenario corrisponde all'esempio medico-storico dello studio sui calcoli renali (Charig et al., 1986): la dimensione del calcolo (Z) influenza sia la scelta del trattamento (X) sia la probabilità di successo (Y), producendo un'inversione nella direzione dell'effetto apparente del trattamento.

Seconda risoluzione: Z è un mediatore. Il meccanismo causale ha la struttura $X \rightarrow Z \rightarrow Y$, dove Z è una variabile intermedia sul percorso causale da X a Y. In questo caso, condizionare su Z blocca il percorso causale da X a Y attraverso Z, producendo una stima dell'effetto diretto (non mediato) invece dell'effetto totale. La distribuzione post-intervento corretta è $P(Y=1 \mid \text{do}(X=1)) = P(Y=1 \mid X=1)$ — la misura marginale non condizionata — che cattura l'effetto totale di X su Y incluso il percorso mediato. La misura stratificata stima l'effetto diretto escludendo la mediazione — una quantità distinta con diverso significato interpretativo. Nel contesto dell'esempio di Berkeley, se il dipartimento scelto è un mediatore tra il genere del candidato e la probabilità di ammissione, allora condizionare per dipartimento produce una stima dell'effetto diretto del genere sull'ammissione al netto del percorso mediato attraverso la scelta del dipartimento — e questa stima, pur essendo calcolata correttamente, risponde a una domanda causale diversa.

Terza risoluzione: Z è un collider. Il meccanismo causale ha la struttura $X \rightarrow Z \leftarrow Y$, dove Z è causato da entrambi X e Y. Il collider bias — noto anche come paradosso di Berkson nella letteratura medica, da Joseph Berkson che lo identificò nel contesto delle statistiche ospedaliere nel 1946 — è il meno intuitivo dei tre meccanismi ma di grandissima rilevanza pratica. Condizionare su un collider apre un percorso di dipendenza tra le sue cause: anche se X e Y sono causalmente indipendenti, condizionare su Z produce una dipendenza artificiosa tra di essi. Nello studio su pazienti ospedalizzati ($Z = \text{ospedalizzato} = 1$), la correlazione negativa tra fumo (X) e cancro al polmone (Y) osservata nel campione condizionato $Z=1$ è un artefatto del bias da collider: sia il fumo sia il cancro aumentano la probabilità di ospedalizzazione, e condizionare su $Z=1$ apre un percorso di dipendenza tra fumo e cancro attraverso l'ospedalizzazione stessa.

La lezione formale della triplice risoluzione è il nucleo più profondo del paradosso: la stessa distribuzione osservazionale $P(X,Y,Z)$ è compatibile con tutti e tre i modelli causali sopra descritti. Non esiste alcuna statistica calcolabile da $P(X,Y,Z)$ — nessun test d'ipotesi, nessun indice di correlazione, nessuna analisi fattoriale, nessun algoritmo di apprendimento automatico che operi sulla distribuzione congiunta — che possa distinguere tra esse. La scelta della misura causalmente corretta dipende dalla struttura causale, e la struttura causale è una proprietà del meccanismo generatore dei dati che non è contenuta nella distribuzione che quel meccanismo produce.

V. La nozione di d-separazione come chiave di lettura del paradosso

V.a Definizione formale e fondamento semantico

Prima di procedere verso la formalizzazione della gerarchia causale, è indispensabile sviluppare in questa sede il criterio di d-separazione nella sua forma rigorosa e nella sua piena profondità teorica, perché esso costituisce non solo la chiave di lettura grafica del paradosso di Simpson nelle sue tre varianti, ma il linguaggio con cui l'intera teoria causale esprime le relazioni tra struttura del modello e proprietà osservazionali. La d-separazione — d da "direzionale", per sottolineare che il blocco dei percorsi dipende dall'orientazione degli archi — fu proposta inizialmente da Thomas Verma e sistematicamente sviluppata da Geiger, Verma e Pearl in una serie di contributi tra il 1987 e il 1990, culminati nel lavoro

fondamentale di Pearl del 1988 *Probabilistic Reasoning in Intelligent Systems*. Il suo significato semantico è di tradurre la struttura qualitativa del DAG — la direzione degli archi, la topologia delle v-strutture — in asserzioni quantitative sulle indipendenze condizionali della distribuzione probabilistica generata dal modello.

La d-separazione opera su percorsi in un DAG G , dove per percorso si intende una sequenza di nodi distinti $V_1, V_2, \dots, V_k$ tali che ogni coppia consecutiva (V_i, V_{i+1}) sia unita da un arco nel grafo non orientato sottostante — senza riguardo alla direzione degli archi. Ogni nodo interno V_i (con $1 < i < k$) può svolgere sul percorso uno di tre ruoli strutturalmente distinti rispetto ai suoi vicini sul percorso: il ruolo di *collider* se entrambi gli archi adiacenti sul percorso sono entranti, cioè se $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$; il ruolo di *catena* (o *mediatore*) se i due archi sono orientati concordemente, cioè se $V_{i-1} \rightarrow V_i \rightarrow V_{i+1}$ oppure $V_{i-1} \leftarrow V_i \leftarrow V_{i+1}$; il ruolo di *biforcazione* (o *causa comune*) se i due archi sono entrambi uscenti, cioè $V_{i-1} \leftarrow V_i \rightarrow V_{i+1}$. La tripartizione è esaustiva: non vi sono altri pattern possibili per un nodo interno di un percorso in un DAG.

Definizione (d-separazione, Pearl 1988). Dato un DAG G e un insieme $C \subseteq V$ di variabili di condizionamento, un percorso π tra A e B in G è *bloccato da* C se almeno una delle seguenti condizioni si verifica per qualche nodo interno V_i di π :

- V_i è un noncollider su π (cioè svolge il ruolo di catena o biforcazione) e $V_i \in C$;
- V_i è un collider su π e né V_i né alcun discendente di V_i appartiene a C , cioè $(\text{Desc}(V_i) \cup \{V_i\}) \cap C = \emptyset$.

Due nodi A e B sono *d-separati* da C in G , scritto $A \perp_d B \mid C$ in G , se ogni percorso tra A e B in G è bloccato da C . Se A e B non sono d-separati da C , si scrive $A \not\perp_d B \mid C$ e si dice che A e B sono *d-connessi* tramite C .

La semantica della definizione non è arbitraria: riflette il comportamento della propagazione dell'informazione in un sistema causale. Un percorso rappresenta un canale potenziale attraverso cui l'informazione su A può influenzare la probabilità di B . I noncollider (catene e biforcazioni) trasmettono l'informazione in modo diretto: se il nodo intermedio è condizionato, il canale è interrotto, come osservare la temperatura nella sequenza sole $\rightarrow$ temperatura $\rightarrow$ sudorazione interrompe la dipendenza tra sole e sudorazione condizionando sulla temperatura. I collider hanno un comportamento opposto e controintuitivo: in assenza di condizionamento il canale è chiuso (le cause di un effetto comune non si "parlano" a meno che non si conosca l'effetto); condizionando sull'effetto comune — o su qualsiasi suo discendente — il canale si apre, creando una dipendenza artificiale tra le cause. Questa asimmetria tra noncollider e collider è il nucleo logico del paradosso di Simpson nel caso collider, e la comprensione delle sue implicazioni è indispensabile per evitare inferenze causali errate nelle analisi osservazionali.

V.b L'algoritmo di Bayes-Ball e la complessità computazionale

La verifica computazionale della d-separazione per una specifica tripla (A, B, C) in un DAG arbitrario non richiede l'enumerazione esplicita di tutti i percorsi — un'operazione potenzialmente esponenziale nel numero di nodi. L'algoritmo di Bayes-Ball di Shachter (1998) risolve il problema in tempo lineare $O(|V| + |E|)$ rispetto al numero di nodi e archi

del grafo, simulando la propagazione di "palline" che obbediscono alle regole di blocco della d-separazione.

L'algoritmo mantiene due insiemi per ogni nodo v : $via-child(v)$, indicante che la pallina ha raggiunto v attraverso un figlio (arco entrante nel percorso verso v), e $via-parent(v)$, indicante che la pallina ha raggiunto v attraverso un genitore (arco uscente nel percorso verso v). Le regole di propagazione sono le seguenti, dove "schedulare" significa aggiungere il nodo alla coda di visita:

Se la pallina arriva a v da un figlio (via-child): se $v \notin C$, schedulare tutti i genitori di v come via-child e tutti i figli di v come via-parent; se $v \in C$, la pallina è bloccata in questa direzione.

Se la pallina arriva a v da un genitore (via-parent): se $v \notin C$, schedulare tutti i figli di v come via-parent; se $v \in C$, schedulare tutti i genitori di v come via-child (il condizionamento apre i collider).

La d-separazione $A \perp_d B \mid C$ è verificata se e soltanto se nessuna pallina partita da A raggiunge B con le regole sopra, con l'inizializzazione di palline di entrambi i tipi (via-child e via-parent) da A . La correttezza dell'algoritmo rispetto alla definizione formale di d-separazione si dimostra per induzione sulla lunghezza dei percorsi, verificando che ogni decisione di propagazione o blocco corrisponde esattamente alle condizioni della definizione.

Una conseguenza pratica rilevante dell'algoritmo è la possibilità di identificare il *Markov blanket* di un nodo A — l'insieme minimo $MB(A)$ tale che $A \perp_d (V \setminus MB(A) \setminus \{A\}) \mid MB(A)$ — in tempo $O(|V| + |E|)$. In un DAG, il Markov blanket di A è esattamente l'unione dei genitori di A , dei figli di A , e degli altri genitori dei figli di A (i "co-genitori"). Questa struttura è fondamentale per gli algoritmi di scoperta causale locale: l'algoritmo MMHC (Max-Min Hill-Climbing) e i suoi discendenti stimano prima il Markov blanket di ogni nodo usando test statistici locali, e usano questa struttura per inizializzare la ricerca globale dello scheletro.

V.c Le tre strutture elementari e le loro implicazioni di d-separazione

Le tre risoluzioni del paradosso di Simpson corrispondono ai tre pattern strutturali elementari che una terza variabile Z può formare con X e Y in un DAG. L'analisi esplicita delle implicazioni di d-separazione per ciascuna struttura costituisce il collegamento formale tra la geometria del paradosso e la teoria generale.

Struttura 1: Il confondente (biforcazione $X \leftarrow Z \rightarrow Y$ con $X \rightarrow Y$). Il grafo ha quattro archi: $Z \rightarrow X$, $Z \rightarrow Y$, $X \rightarrow Y$. I percorsi tra X e Y sono due: il percorso diretto $\pi_1 = (X \rightarrow Y)$ e il percorso di backdoor $\pi_2 = (X \leftarrow Z \rightarrow Y)$.

Sul percorso π_1 , non vi sono nodi interni: il percorso è sempre attivo per ogni insieme di condizionamento C che non contenga X o Y .

Sul percorso π_2 , il nodo interno è Z , che ha il ruolo di noncollider di tipo biforcazione ($X \leftarrow Z \rightarrow Y$). Le regole della d-separazione danno: — π_2 è bloccato da C se e solo se $Z \in C$.

Senza condizionamento ($C = \emptyset$): π_2 è aperto, e la d-connessione di X e Y riflette sia l'effetto causale diretto (attraverso π_1) sia la dipendenza spuria (attraverso π_2). La correlazione osservazionale $\rho_{\{XY\}}$ cattura entrambi.

Condizionando su Z ($C = \{Z\}$): π_2 è bloccato. La d-connessione residua è solo quella di π_1 , che riflette l'effetto causale diretto $X \rightarrow Y$. Il coefficiente di regressione parziale $\beta_{\{Y|X, Z=z\}}$ stima l'effetto causale diretto, al netto del confondimento.

L'aggiustamento per Z — la formula di backdoor nella sua versione elementare — è pertanto:

$$P(Y \mid \text{do}(X = x)) = \sum_z P(Y \mid X = x, Z = z) \cdot P(Z = z)$$

che calcola la media pesata dei coefficienti di regressione parziale sulle coorti di Z, eliminando la dipendenza spuria introdotta dal percorso di backdoor.

Struttura 2: Il mediatore (catena $X \rightarrow Z \rightarrow Y$ con effetto diretto $X \rightarrow Y$). Il grafo ha archi $X \rightarrow Z$, $Z \rightarrow Y$, $X \rightarrow Y$. I percorsi tra X e Y sono: $\pi_1 = (X \rightarrow Y)$ e $\pi_2 = (X \rightarrow Z \rightarrow Y)$.

Sul percorso π_2 , il nodo interno è Z, con ruolo di noncollider di tipo catena. Le implicazioni di d-separazione: — Senza condizionamento ($C = \emptyset$): π_2 è aperto, e la correlazione cattura sia l'effetto diretto sia l'effetto mediato attraverso Z. — Condizionando su Z ($C = \{Z\}$): π_2 è bloccato. La correlazione residua riflette solo l'effetto diretto $X \rightarrow Y$, al netto della mediazione.

In questo caso, la domanda causale determina la corretta procedura: se si vuole stimare l'effetto *totale* di X su Y — inclusa la mediazione attraverso Z — non si deve condizionare su Z; se si vuole stimare l'effetto *diretto* (non mediato) di X su Y, si condiziona su Z. Il paradosso di Simpson nel caso mediatore nasce dall'inversione di questa logica: aggregare i dati equivale a non condizionare su Z; stratificare per Z equivale a condizionare su Z. La domanda di quale delle due analisi sia "corretta" non ha risposta statistica: dipende dall'obiettivo causale dell'analisi.

La decomposizione effetto totale / effetto diretto / effetto indiretto si formalizza come:

$$\text{ATE}(X \rightarrow Y) = \text{ADE}(X \rightarrow Y) + \text{AIE}(X \rightarrow Z \rightarrow Y)$$

dove ADE è l'Average Direct Effect e AIE è l'Average Indirect Effect (o Average Mediated Effect). Nel modello lineare gaussiano senza interazioni:

$$\beta_{\text{total}} = \beta_{\text{direct}} + \beta_{\{XZ\}} \cdot \beta_{\{ZY\}}$$

dove β_{direct} è il coefficiente di regressione di Y su X con Z inclusa nel modello, e $\beta_{\{XZ\}} \cdot \beta_{\{ZY\}}$ è il prodotto dei coefficienti sul percorso mediato. Nel caso non lineare o con interazioni, la decomposizione richiede il formalismo dell'effetto diretto naturale (Natural Direct Effect) e dell'effetto indiretto naturale (Natural Indirect Effect) di Pearl, che fa uso di interventi imbricati $\text{do}(X = x, Z = Z_{\{x'\}})$: fissare X a x mentre Z assume il valore che avrebbe se X fosse stato x' , una quantità controfattuale genuina di livello-3 della gerarchia.

Struttura 3: Il collider ($X \rightarrow Z \leftarrow Y$ senza arco diretto $X \rightarrow Y$). Il grafo ha archi $X \rightarrow Z$ e $Y \rightarrow Z$, con X e Y altrimenti indipendenti. I percorsi tra X e Y sono uno solo: $\pi = (X \rightarrow Z \leftarrow Y)$.

Sul percorso π , il nodo interno è Z , con ruolo di collider. Le regole della d-separazione danno: — Senza condizionamento ($C = \emptyset$): il collider Z blocca naturalmente il percorso. X e Y sono d-separati: $A \perp_d B \mid \emptyset$. La correlazione osservazionale $P(X, Y) = P(X) \cdot P(Y)$: X e Y sono statisticamente indipendenti. — Condizionando su Z ($C = \{Z\}$): il collider Z è attivato. Il percorso π diventa attivo. X e Y diventano dipendenti condizionalmente a Z , anche se marginalmente indipendenti: $A \not\perp_d B \mid \{Z\}$.

Questo è il meccanismo fondamentale del *collider bias*, o *selection bias* nella terminologia epidemiologica, o *Berkson's paradox* nella nomenclatura clinica: selezionare su Z (l'effetto comune) induce una correlazione artificiale tra X e Y nella sottopopolazione selezionata, anche quando X e Y sono causalmente indipendenti. Il paradosso di Simpson nel caso collider si manifesta quando Z è la variabile di stratificazione: l'analisi stratificata produce, in ciascun strato, una correlazione tra X e Y che non riflette alcuna relazione causale, ma è artefatto della stratificazione stessa. Nella struttura confondente il condizionamento su Z elimina la correlazione spuria; nella struttura collider il condizionamento su Z la introduce — la distinzione è diametralmente opposta, e la distribuzione osservazionale non permette di discriminare tra le due strutture.

La dipendenza condizionale introdotta dal collider ha una struttura precisa nel modello lineare gaussiano. Siano $X \sim N(0, \sigma^2_X)$ e $Y \sim N(0, \sigma^2_Y)$ indipendenti, e sia $Z = \alpha_X \cdot X + \alpha_Y \cdot Y + \epsilon_Z$ con $\epsilon_Z \sim N(0, \sigma^2_{\epsilon})$. La correlazione condizionale di X e Y dato $Z = z$ si calcola come:

$$\rho(X, Y \mid Z = z) = -\alpha_X \cdot \alpha_Y \cdot \sigma_X \cdot \sigma_Y / \sqrt{(\alpha_X^2 \cdot \sigma_X^2 + \sigma_{\epsilon}^2)} / \sqrt{(\alpha_Y^2 \cdot \sigma_Y^2 + \sigma_{\epsilon}^2)}$$

che è negativa se α_X e α_Y hanno lo stesso segno. Questo è il *explaining away* (o *discounting*): osservare $Z = z$ con valore elevato suggerisce che almeno una delle cause contribuisce sostanzialmente; se X è elevato, la necessità di Y per spiegare Z si riduce, creando la correlazione negativa. Il fenomeno è strutturale: persiste per qualsiasi dimensione del campione e non è dovuto a errori di stima.

V.d La simmetria del paradosso e l'indistinguibilità strutturale

La connessione tra la tripartizione strutturale (confondente, mediatore, collider) e le tre manifestazioni del paradosso di Simpson è ora formalizzabile con precisione. In tutte e tre le strutture, la distribuzione marginale $P(X, Z)$ e la distribuzione marginale $P(Y, Z)$ possono essere identiche — e quindi indistinguibili statisticamente. La differenza è nella distribuzione condizionale di Y dato X e Z , e specificamente nel segno e nella magnitudo della dipendenza residua $Y \mid X, Z$ rispetto alla dipendenza marginale $Y \mid X$.

Più precisamente: nelle strutture confondente e mediatore, il coefficiente di regressione parziale $\beta_{Y|X, Z=z}$ differisce dal coefficiente marginale $\beta_{Y|X}$ nella stessa direzione (per il confondente nella direzione del bias di confondimento; per il mediatore nella direzione dell'effetto diretto), ma non necessariamente in modo da invertirne il segno. Il paradosso si

manifesta nel caso confondente quando il confondimento è sufficientemente forte da dominare l'effetto causale diretto: la correlazione marginale è determinata principalmente dal percorso di backdoor, mentre la correlazione parziale riflette solo l'effetto diretto, che può essere di segno opposto.

Nella struttura collider, la situazione è strutturalmente diversa: la correlazione marginale è esattamente zero (indipendenza) mentre la correlazione parziale è non nulla — il condizionamento la crea ex nihilo. Il paradosso di Simpson nel caso collider è quindi più marcato nella sua inversione: si passa da $\rho = 0$ a $\rho \neq 0$ con un segno determinato dalla struttura del collider. Ma è anche più pericoloso epistemicamente: mentre nel caso confondente un'analisi stratificata produce stime della relazione causale $X \rightarrow Y$ (parzialmente corrette), nel caso collider la stessa analisi stratificata produce stime che non hanno alcun contenuto causale relativo all'effetto di X su Y , essendo completamente determinate dall'artefatto di selezione.

La dimostrazione formale dell'indistinguibilità delle tre strutture dalla sola distribuzione osservazionale è la seguente. Si consideri la distribuzione gaussiana trivariata (X, Y, Z) con matrice di varianza:

$$\Sigma = \begin{vmatrix} \sigma^2_X & \rho_{XY} \cdot \sigma_X \cdot \sigma_Y & \rho_{XZ} \cdot \sigma_X \cdot \sigma_Z \\ \rho_{XY} \cdot \sigma_X \cdot \sigma_Y & \sigma^2_Y & \rho_{YZ} \cdot \sigma_Y \cdot \sigma_Z \\ \rho_{XZ} \cdot \sigma_X \cdot \sigma_Z & \rho_{YZ} \cdot \sigma_Y \cdot \sigma_Z & \sigma^2_Z \end{vmatrix}$$

Per qualsiasi Σ positiva definita con $\rho_{XY} \neq 0$, $\rho_{XZ} \neq 0$, $\rho_{YZ} \neq 0$, esistono almeno due SCM con grafi distinti G_1 (confondente: $X \leftarrow Z \rightarrow Y$, $X \rightarrow Y$) e G_2 (mediatore: $X \rightarrow Z \rightarrow Y$, $X \rightarrow Y$) che producono la medesima Σ . Questo perché la matrice Σ ha sei gradi di libertà (σ_X , σ_Y , σ_Z , ρ_{XY} , ρ_{XZ} , ρ_{YZ}) e il modello lineare gaussiano con tre variabili ha in generale sufficienti parametri liberi per adattarsi a qualsiasi Σ positiva definita con entrambe le strutture di grafo. La differenza emerge soltanto nella distribuzione post-intervento: $P_{\{G_1\}}(Y \mid \text{do}(X = x)) \neq P_{\{G_2\}}(Y \mid \text{do}(X = x))$ in generale, anche quando $P_{\{G_1\}}(X, Y, Z) = P_{\{G_2\}}(X, Y, Z)$. Questo è precisamente il contenuto operativo del teorema di incommensurabilità applicato al caso trivariato.

V.e La condizione di Markov causale e la fedeltà come assiomi del formalismo

Il collegamento tra d-separazione nel DAG e indipendenze condizionali nella distribuzione è fornito da due assiomi che costituiscono il fondamento del formalismo:

La *condizione di Markov causale* afferma che ogni variabile endogena è condizionalmente indipendente dai propri non-discendenti, dati i propri genitori. Formalmente, per ogni $V_i \in V$:

$$V_i \perp (V \setminus \{V_i\} \setminus \text{Desc}(V_i)) \mid \text{PA}(V_i)$$

dove $\text{Desc}(V_i)$ è l'insieme dei discendenti di V_i nel DAG. La condizione di Markov è soddisfatta da qualsiasi distribuzione generata da un SCM con variabili esogene mutuamente indipendenti — una conseguenza diretta della struttura fattorizzata del modello generativo. Essa implica la *condizione di fattorizzazione globale*:

$$P(V_1, \dots, V) = \prod_i P(V_i \mid \text{PA}(V_i))$$

che permette di leggere le indipendenze condizionali implicate dal DAG direttamente dalla struttura di d-separazione, tramite il teorema di d-separazione e indipendenza condizionale:

$$A \perp_d B \mid C \text{ in } G \Rightarrow A \perp B \mid C \text{ in } P$$

La condizione di Markov è la versione locale del principio reichenbachiano della causa comune: ogni dipendenza residua tra V_i e i suoi non-discendenti dopo il condizionamento sui genitori è una dipendenza non spiegata da cause esplicite nel modello, il che segnala incompletezza della specificazione causale.

La *condizione di fedeltà* è la condizione complementare: afferma che tutte le indipendenze presenti in P sono implicate dalla struttura di d-separazione del DAG — non vi sono indipendenze "accidentali" dovute a cancellazioni parametriche:

$$A \perp B \mid C \text{ in } P \Rightarrow A \perp_d B \mid C \text{ in } G$$

Insieme, Markov e fedeltà stabiliscono la corrispondenza biunivoca tra la struttura di d-separazione nel DAG e la struttura di indipendenza condizionale nella distribuzione, fondamento della scoperta causale osservazionale. La fedeltà è violabile — come discusso nella Parte Terza — ma è generica nel senso della misura di Lebesgue: le violazioni si verificano su insiemi di parametri di misura nulla nello spazio dei modelli.

Il **teorema di d-separazione e indipendenza condizionale** afferma in modo più preciso che la condizione di Markov implica la fattorizzazione globale della distribuzione secondo la struttura del DAG, dalla quale seguono tutte le indipendenze condizionali implicite dal criterio di d-separazione. La dimostrazione usa il teorema di Geiger, Verma e Pearl (1990) sulla completezza della d-separazione rispetto alla condizione di Markov: sotto la condizione di Markov, l'insieme delle indipendenze condizionali implicite dalla d-separazione è esattamente l'insieme delle indipendenze condizionali valide per tutte le distribuzioni compatibili con il grafo (cioè per tutte le scelte dei parametri delle equazioni strutturali compatibili con il DAG). In altri termini, la d-separazione cattura tutte e sole le indipendenze strutturali — quelle che valgono indipendentemente dai parametri specifici.

V.f Conseguenze formali per le tre risoluzioni del paradosso

Alla luce dell'analisi delle tre strutture e della teoria della d-separazione, è ora possibile dare alle tre risoluzioni del paradosso di Simpson lo statuto formale che meritano, superando la presentazione convenzionale in termini di direzioni degli archi e chiarendo in cosa consiste precisamente la scelta tra aggregazione e stratificazione.

La risoluzione corretta del paradosso non è determinata dalla distribuzione $P(X, Y, Z)$ — che è la stessa nelle tre strutture nel caso di equivalenza osservazionale — ma dalla struttura causale del DAG, che determina se Z è un confondente (da aggiustare), un mediatore (da non aggiustare se si vuole l'effetto totale, da aggiustare se si vuole l'effetto diretto), o un collider (da non aggiustare sotto nessuna circostanza). Questa determinazione richiede, necessariamente, informazioni che vanno al di là della distribuzione osservazionale: richiede

teoria del dominio, esperimenti di intervento, o almeno assunzioni funzionali aggiuntive (non-gaussianità, addittività del rumore) che permettano l'identificazione della struttura.

Il paradosso di Simpson non è paradossale nella struttura matematica — la distribuzione si comporta esattamente come la teoria delle probabilità prevede — ma nella pretesa epistemica di inferire la struttura causale dai soli dati osservazionali. Questa pretesa è strutturalmente impossibile, e la sua impossibilità è enunciata con precisione formale nel teorema di incommensurabilità della sezione VI e nella teoria dell'equivalenza di Markov della Parte Terza. Il paradosso è il monumento più elementare a questa impossibilità: con sole tre variabili e due osservazioni (la correlazione marginale e la correlazione parziale), la struttura causale è radicalmente sottodeterminata. Con più variabili, il problema si moltiplica proporzionalmente alla dimensione dell'ambiguità nelle classi di equivalenza di Markov, la cui geometria la Parte Terza tratta sistematicamente.

VI. Il confine formale tra osservazione e intervento: il teorema di incommensurabilità

Il paradosso di Simpson è la manifestazione più elementare di una barriera logica più profonda, che Pearl ha formalizzato nella distinzione tra probabilità condizionale osservazionale e probabilità post-intervento. Questa distinzione può essere enunciata nella sua forma più netta come teorema di incommensurabilità tra il livello-1 e il livello-2 della gerarchia causale.

Per enunciare il teorema con precisione è necessario richiamare la struttura dei modelli causali strutturali (SCM). Un modello causale strutturale M è una quadrupla $\langle U, V, \mathcal{F}, P(U) \rangle$ dove $U = \{U_1, \dots, U_n\}$ è l'insieme delle variabili esogene con distribuzione fattorizzata $P(U) = \prod_i P(U_i)$, $V = \{V_1, \dots, V_m\}$ è l'insieme delle variabili endogene osservabili, e $\mathcal{F} = \{f_1, \dots, f_m\}$ è l'insieme delle equazioni strutturali con $V_i := f_i(PA(V_i), U_i)$, dove $PA(V_i) \subseteq V \setminus \{V_i\}$ sono i genitori causali di V_i . Il sistema di equazioni strutturali ammette un ordinamento topologico — la condizione di aciclicità del grafo indotto.

L'operatore di intervento $do(V_i = v_i)$ agisce su M producendo il modello mutilato $M_{\leftarrow \{V_i \leftarrow v_i\}}$, ottenuto sostituendo l'equazione strutturale di V_i con la costante $V_i := v_i$, lasciando inalterate tutte le altre equazioni. Nel grafo G indotto da M , questa operazione corrisponde alla rimozione di tutti gli archi entranti in V_i , producendo il grafo parzialmente reciso $G_{\leftarrow \{V_i\}} = \{V_i\}$. La distribuzione $P(Y \mid do(X=x))$ è la distribuzione di Y nel modello mutilato $M_{\leftarrow \{X \leftarrow x\}}$.

Teorema di incommensurabilità (Pearl, 1995; forma estesa). Esistono due modelli causali strutturali M_1 e M_2 con il medesimo grafo G (o con grafi distinti G_1 e G_2 appartenenti alla stessa classe di equivalenza di Markov) tali che per ogni $\epsilon > 0$:

$$d_{TV}(P_{\leftarrow \{M_1\}}(V), P_{\leftarrow \{M_2\}}(V)) < \epsilon$$

e contemporaneamente la variazione totale tra le distribuzioni post-intervento è non trascurabile per almeno un intervento (X, x) :

$$d_{TV}(P_{\leftarrow \{M_1\}}(Y \mid do(X=x)), P_{\leftarrow \{M_2\}}(Y \mid do(X=x))) \geq \delta > 0$$

indipendentemente da ϵ . In altri termini: per ogni grado di approssimazione ϵ dell'identità osservazionale, esistono modelli osservazionalmente indistinguibili che producono distribuzioni post-intervento arbitrariamente distinte.

La famiglia di modelli che prova questo teorema è quella dei modelli lineari bivariati con variabile latente. Sia $U \sim N(0, 1)$ la variabile latente condivisa, e siano $\epsilon_X \sim N(0, \sigma^2_X)$ ed $\epsilon_Y \sim N(0, \sigma^2_Y)$ variabili di errore mutuamente indipendenti e indipendenti da U . La famiglia parametrica è:

$$X := \alpha \cdot U + \epsilon_X$$

$$Y := \beta \cdot X + \gamma \cdot U + \epsilon_Y$$

I parametri liberi sono $\theta = (\alpha, \beta, \gamma, \sigma^2_X, \sigma^2_Y)$. Le quantità osservabili sono:

$$\text{Var}(X) = \alpha^2 + \sigma^2_X$$

$$\text{Cov}(X, Y) = \beta \cdot \text{Var}(X) + \alpha \cdot \gamma = \beta(\alpha^2 + \sigma^2_X) + \alpha\gamma$$

$$\text{Var}(Y) = (\beta\alpha + \gamma)^2 + \beta^2 \cdot \sigma^2_X + \sigma^2_Y$$

Il coefficiente di regressione OLS di Y su X è:

$$b_{\text{OLS}} = \text{Cov}(X, Y) / \text{Var}(X) = \beta + \alpha\gamma / (\alpha^2 + \sigma^2_X)$$

Il secondo termine è il bias di confondimento, che si annulla se e solo se $\alpha = 0$ (U non influenza X) o $\gamma = 0$ (U non influenza Y). Nel modello mutilato $M_{\{X \leftarrow x\}}$, l'equazione di X viene sostituita dalla costante $X := x$, e la distribuzione di Y diventa:

$$Y = \beta \cdot x + \gamma \cdot U + \epsilon_Y \sim N(\beta \cdot x, \gamma^2 + \sigma^2_Y)$$

$$\text{dunque } \mathbb{E}[Y \mid \text{do}(X=x)] = \beta \cdot x.$$

Consideriamo ora una famiglia di modelli indicizzata dal parametro $c \in (-\sqrt{2}, \sqrt{2})$. Si fissino $\Sigma_X = 2$, $\Sigma_{\{XY\}} = 1$, $\Sigma_Y = 3$, e si consideri per ogni c il modello $M(c)$ con parametri:

$$\alpha = 1, \sigma^2_X = 1, \beta(c) = c, \gamma(c) = 1 - c, \sigma^2_Y(c) = 2 - c^2$$

Verifica di osservazionale equivalenza. $\text{Var}(X) = 1 + 1 = 2$ ✓. $\text{Cov}(X, Y) = c \cdot 2 + 1 \cdot (1 - c) = 2c + 1 - c = c + 1$... attendere: si ricalcola con la formula corretta:

$$\text{Cov}(X, Y) = \beta \cdot \text{Var}(X) + \alpha \cdot \gamma = c \cdot 2 + 1 \cdot (1 - c) = 2c + 1 - c = c + 1.$$

Poiché questa deve valere 1 per ogni c , è necessario correggere la costruzione. Si fissano invece $\alpha = 1$, $\sigma^2_X = 1$, e si richiede $\text{Cov}(X, Y) = 1$ con $\beta(c) = c$ e $\gamma(c)$ da determinare:

$$\text{Cov}(X, Y) = c \cdot 2 + 1 \cdot \gamma(c) = 1 \Rightarrow \gamma(c) = 1 - 2c$$

Si richiede $\text{Var}(Y) = 3$:

$$\text{Var}(Y) = (c \cdot 1 + \gamma(c))^2 + c^2 \cdot 1 + \sigma^2_Y(c) = (c + 1 - 2c)^2 + c^2 + \sigma^2_Y(c) = (1-c)^2 + c^2 + \sigma^2_Y(c)$$

Quindi $\sigma^2_Y(c) = 3 - (1-c)^2 - c^2 = 3 - 1 + 2c - c^2 - c^2 = 2 + 2c - 2c^2$. Affinché $\sigma^2_Y(c) > 0$ si richiede $2 + 2c - 2c^2 > 0$, ovvero $c^2 - c - 1 < 0$, cioè $c \in ((1-\sqrt{5})/2, (1+\sqrt{5})/2) \approx (-0.618, 1.618)$.

All'interno di questo intervallo, il modello $M(c)$ con parametri $\alpha=1$, $\sigma^2_X=1$, $\beta=c$, $\gamma=1-2c$, $\sigma^2_Y=2+2c-2c^2$ è osservazionalmente equivalente a $M(0)$ (con $\beta=0$, $\gamma=1$, $\sigma^2_Y=2$) ma produce l'effetto causale:

$$\mathbb{E}[Y \mid \text{do}(X=x)] = \beta(c) \cdot x = c \cdot x$$

che varia continuamente con c , cambiando di segno a $c=0$ e coprendo un intervallo che include sia valori positivi sia negativi. In particolare $M(c_1)$ e $M(c_2)$ per $c_1 \neq c_2$ sono osservazionalmente identici ma interventisticamente distinti — esattamente la struttura richiesta dal teorema di incommensurabilità. $\square$

Questo risultato — che non è specifico dei modelli lineari gaussiani ma si generalizza a qualsiasi famiglia di SCM con confondimento latente — afferma l'impossibilità di qualsiasi mappa calcolabile da distribuzioni osservazionali a distribuzioni interventistiche che sia universalmente corretta. Non si tratta di un problema di precisione computazionale o di dimensione del campione: anche con la distribuzione osservazionale esatta $P(V)$, la distribuzione post-intervento non è determinata senza conoscenza della struttura causale del grafo.

VII. La gerarchia causale nella sua forma sistematica

Il paradosso di Simpson è il caso paradigmatico del confine tra il primo e il secondo livello di una gerarchia più articolata che Pearl ha denominato Ladder of Causation — la Scala della Causazione. Questa gerarchia classifica la conoscenza causale in tre livelli di potere esplicativo crescente e di accessibilità osservazionale decrescente.

Il livello-1, quello dell'associazione, comprende tutto ciò che può essere calcolato dalla distribuzione di probabilità osservazionale $P(V)$. Le domande di livello-1 hanno la forma: quale è la probabilità di Y dato che si osserva $X=x$? Come si modifica la distribuzione di Y al variare del valore osservato di X ? Quali variabili sono informative su Y nella distribuzione storica? Il formalismo è la teoria standard della probabilità condizionale: $P(Y|X=x)$, $\mathbb{E}[Y|X=x]$, $\text{Corr}(X,Y)$ e così via. Ogni algoritmo di apprendimento automatico addestrato su distribuzioni fisse — le reti neurali profonde, i transformer, i modelli di linguaggio — opera strutturalmente a questo livello, approssimando funzioni della distribuzione osservazionale.

Il livello-2, quello dell'intervento, comprende le domande della forma: quale sarebbe la distribuzione di Y se si intervenisse per fissare il valore di X a x , indipendentemente dalle cause naturali di X ? Il formalismo è l'operatore $\text{do}(\cdot)$: $P(Y \mid \text{do}(X=x))$ è la distribuzione di Y nel modello causale mutilato $M_{\{X \leftarrow x\}}$. Queste domande riguardano gli esperimenti attivi, le politiche deliberate, le azioni dei decisori. La differenza fondamentale tra livello-1 e

livello-2 si manifesta quando il trattamento X è confuso con altre cause di Y : la relazione osservazionale $P(Y|X=x)$ incorpora sia l'effetto causale sia il confondimento, mentre $P(Y|\text{do}(X=x))$ isola l'effetto causale puro.

Il livello-3, quello del controfattuale, comprende le domande della forma: quale sarebbe stato il valore di Y per questo specifico individuo se X fosse stato diverso da ciò che è stato nella realtà? Il formalismo richiede le equazioni strutturali nella loro forma funzionale completa: il controfattuale $Y_{\{X \leftarrow x\}}(u) = f_Y(\text{PA}(Y)$ nel modello $M_{\{X \leftarrow x\}}$ con $U=u$) è il valore di Y per l'individuo u nel mondo ipotetico in cui X è stato fissato a x per intervento. Queste domande riguardano la responsabilità legale e morale, la spiegazione di singoli eventi, la medicina personalizzata. La domanda "questo paziente avrebbe guarito se avesse ricevuto il farmaco?" — per un paziente che non lo ha ricevuto — è una domanda di livello-3: richiede di ragionare sul controfattuale per quell'individuo specifico, condizionato sulle sue caratteristiche osservate.

La strettezza della gerarchia — il fatto che nessun livello inferiore consente di rispondere a domande del livello superiore senza assunzioni aggiuntive — è il teorema centrale delle Parti Seconda e Terza. La transizione dal livello-1 al livello-2 richiede ipotesi sulla struttura del grafo causale, non verificabili da $P(V)$ senza esperimenti di intervento. La transizione dal livello-2 al livello-3 richiede la conoscenza della struttura di dipendenza congiunta delle variabili esogene U , non ricostruibile nemmeno dalla distribuzione interventistica completa — per ragioni che dipendono dall'impossibilità di identificare i meccanismi di generazione degli errori dall'esterno del sistema.

Formalmente, le tre classi di funzioni corrispondenti ai tre livelli formano una gerarchia inclusiva stretta:

$$\mathcal{L}_1 \subset \mathcal{L}_2 \subset \mathcal{L}_3$$

dove $\mathcal{L}_1$ è la classe di tutte le funzioni calcolabili da $P(V)$, $\mathcal{L}_2$ la classe di tutte le funzioni calcolabili da $\{P(V | \text{do}(X=x)) : X \subseteq V, x \in \mathcal{X}\}$ e $\mathcal{L}_3$ la classe di tutte le funzioni calcolabili dalla famiglia completa di distribuzioni controfattuali. Le inclusioni sono strette: esistono funzioni di $\mathcal{L}_2$ non calcolabili da alcun elemento di $\mathcal{L}_1$ (il teorema di incommensurabilità), ed esistono funzioni di $\mathcal{L}_3$ non calcolabili da alcun elemento di $\mathcal{L}_2$ (il teorema di non-identificabilità dei controfattuali della Parte Seconda).

VIII. Il paradosso come monumento al confine

Torniamo al paradosso di Simpson per precisarne il ruolo nell'architettura dell'argomentazione con la chiarezza che merita. La sua funzione in questa monografia non è quella di un esempio motivante che viene poi abbandonato: è quella di un filo conduttore che accompagna l'intera trattazione formale, incarnando nel caso più semplice possibile ogni aspetto generale del confine che stiamo esplorando.

Il paradosso mostra — nel caso di sole tre variabili binarie — che la stessa distribuzione osservazionale è compatibile con interpretazioni causali opposte. Non si tratta di ambiguità nell'interpretazione di evidenze deboli o di dati rumorosi: si tratta di incompatibilità logica tra interpretazioni egualmente consistenti con i dati, indipendentemente dalla dimensione del

campione. Un campione infinito della distribuzione dell'esempio della sezione II lascerebbe irrisolta la domanda causale allo stesso modo di un campione finito: perché la risposta alla domanda causale non è nella distribuzione ma nel meccanismo che la genera — una proprietà che abita un livello epistemico distinto e non accessibile per via meramente osservazionale.

Il paradosso illustra anche il meccanismo fondamentale attraverso cui i sistemi di intelligenza artificiale addestrati su dati osservazionali possono produrre raccomandazioni causalmente errate: essi ottimizzano $P(Y|X=x)$ — la probabilità condizionale osservazionale — anziché $P(Y|\text{do}(X=x))$ — la distribuzione post-intervento. Quando il confondimento è presente, le due quantità differiscono, e la differenza si manifesta come errore sistematico nelle raccomandazioni interventistiche. La struttura del bias è sempre la stessa: il sistema confonde la correlazione con la causa, assegnando all'azione X il merito (o il demerito) di Y che invece appartiene al confondimento U . Il paradosso di Simpson è la forma più elementare di questo errore, e la sua analisi formale è la chiave di comprensione degli errori molto più complessi che i sistemi di intelligenza artificiale commettono nei contesti decisionali reali.

Il paradosso costituisce infine il banco di prova su cui ogni proposta di soluzione algoritmica del problema causale deve essere valutata: un sistema che non risolve correttamente il paradosso di Simpson nei casi in cui la struttura causale è nota non offre alcuna garanzia per i casi in cui la struttura non è nota. E un sistema che risolve il paradosso solo perché gli viene fornita la struttura causale non ha risolto il problema, lo ha spostato: la domanda diventa dove quella struttura viene acquisita, e la risposta — che non può essere "dai dati osservazionali senza assunzioni aggiuntive", pena la circolarità — è precisamente il tema di tutta la monografia.

Parte Seconda: La Gerarchia Causale come Sistema di Teoremi di Impossibilità

I. La strettezza della gerarchia: formulazione precisa del problema

L'intuizione che associazione, intervento e controfattuale formino livelli distinti di conoscenza è presente, in forme diverse, in quasi ogni trattazione della causalità a partire da Hume. Ciò che manca, nella maggior parte di queste trattazioni, è una dimostrazione rigorosa che la gerarchia sia *stretta* nel senso tecnico preciso: che non esista alcuna mappa computabile dal livello-1 al livello-2 o dal livello-2 al livello-3 che funzioni universalmente — cioè per ogni modello causale strutturale in una classe sufficientemente ampia, non soltanto per classi speciali o sotto assunzioni restrittive non esplicitate. Il programma di questa Parte è costruire tale dimostrazione per stadi successivi, partendo dal risultato più elementare e procedendo verso i teoremi di impossibilità più profondi, preservando in ogni passaggio la chiarezza sulle premesse effettivamente necessarie.

Prima di entrare nelle dimostrazioni è essenziale fissare i termini in modo rigoroso. Un modello causale strutturale (SCM) $M = \langle U, V, \mathcal{F}, P(U) \rangle$ è definito dalle seguenti componenti.

$U = \{U_1, \dots, U\}$ è l'insieme delle variabili esogene, con distribuzione congiunta $P(U) = \prod_i P(U_i)$ — l'assunzione di indipendenza mutua delle esogene è essenziale per la teoria e va mantenuta esplicita, poiché la sua violazione produce la struttura di dipendenza tra esogene che rende i controfattuali non identificabili, come mostrerà la Sezione VII. $V = \{V_1, \dots, V\}$ è l'insieme delle variabili endogene, osservabili in linea di principio. $\mathcal{F} = \{f_1, \dots, f\}$ è l'insieme delle equazioni strutturali, dove ogni equazione ha la forma $V_i := f_i(\text{PA}(V_i), U_i)$ con $\text{PA}(V_i) \subseteq V \setminus \{V_i\}$ i genitori causali di V_i — cioè le variabili endogene che appaiono come argomenti diretti nella funzione strutturale di V_i .

L'SCM M induce un grafo diretto aciclico $G(M)$ i cui nodi sono le variabili endogene V e i cui archi sono determinati dalle dipendenze nelle equazioni strutturali: vi è un arco $V \rightarrow V_i$ se e solo se $V \in \text{PA}(V_i)$. L'aciclicità è garantita dall'assunzione che il sistema di equazioni ammetta un ordinamento topologico $V_{\pi(1)}, \dots, V_{\pi(n)}$ tale che ogni variabile V_i dipende strutturalmente solo da variabili precedenti nell'ordinamento. L'ordinamento topologico non è necessariamente unico, ma la sua esistenza è equivalente all'aciclicità del DAG.

L'SCM M induce una distribuzione osservazionale $P_M(V)$ tramite la composizione delle equazioni strutturali nell'ordinamento topologico: si campiona $P(U)$, si propagano i valori delle esogene attraverso le equazioni nella sequenza $V_{\pi(1)}, \dots, V_{\pi(n)}$, ottenendo i valori delle endogene come funzione deterministica delle esogene dato il sistema di equazioni.

L'operatore di intervento $\text{do}(V_i = v_i)$ agisce producendo il modello mutilato $M_{\{V_i \leftarrow v_i\}}$: si sostituisce l'equazione strutturale f_i con la costante $V_i := v_i$, e si lascia inalterata ogni altra equazione f_j con $j \neq i$. Nel grafo $G(M)$, questa operazione corrisponde alla rimozione di tutti gli archi entranti in V_i — il nodo i diventa una sorgente nel grafo mutilato. La distribuzione $P(Y \mid \text{do}(X=x))$ è la distribuzione marginale di Y nel modello mutilato $M_{\{X \leftarrow x\}}$.

Questa formalizzazione permette di enunciare con precisione la domanda centrale: data la distribuzione osservazionale $P_M(V)$ e il grafo $G(M)$, la distribuzione interventistica $P_M(Y \mid \text{do}(X=x))$ è univocamente determinata? Esiste una funzione calcolabile $\mathbb{F}$ tale che $P_M(Y \mid \text{do}(X=x)) = \mathbb{F}(P_M(V), G(M))$ per ogni SCM M compatibile con $G(M)$ e con $P_M(V)$?

II. Il teorema fondamentale di non-identificabilità

Teorema (non-identificabilità in presenza di confondimento latente). Siano dati un grafo G con almeno una coppia di nodi (X, Y) tali che X sia una causa diretta di Y e vi sia un confounder latente U — un nodo non osservabile con archi verso sia X sia Y . Allora per ogni distribuzione osservazionale $P(X, Y)$ esiste una famiglia $\{M_\beta\}$ parametrizzata da un parametro reale β appartenente a un intervallo aperto non degenere, tale che:

$$P_{\{M_\beta\}}(X, Y) = P(X, Y) \text{ per ogni } \beta \text{ nell'intervallo}$$

e simultaneamente l'effetto causale $\mathbb{E}_{\{M_\beta\}}[Y \mid \text{do}(X=x)] = \beta \cdot x + \kappa$ varia con β al variare del parametro, per qualche costante κ .

Dimostrazione per costruzione nel modello lineare gaussiano. Si considerano tre quantità osservazionali target: $\Sigma_X = \text{Var}(X)$, $\Sigma_{XY} = \text{Cov}(X, Y)$, $\Sigma_Y = \text{Var}(Y)$, con $\Sigma_X > 0$ e

$\Sigma_Y > 0$. Si parametrizza la famiglia di modelli fissando $\alpha > 0$ e facendo variare il coefficiente causale β :

$$X := \alpha \cdot U + \epsilon_X, \text{ con } U \sim N(0,1) \text{ e } \epsilon_X \sim N(0, \Sigma_X - \alpha^2) \text{ per } \alpha^2 < \Sigma_X$$

$$Y := \beta \cdot X + \gamma(\beta) \cdot U + \epsilon_Y(\beta)$$

dove $\gamma(\beta)$ e $\sigma^2_Y(\beta)$ sono determinati dai vincoli di compatibilità osservazionale:

$$\text{Cov}(X, Y) = \beta \cdot \text{Var}(X) + \alpha \cdot \gamma(\beta) \Rightarrow \gamma(\beta) = (\Sigma_{\{XY\}} - \beta \cdot \Sigma_X) / \alpha$$

$$\text{Var}(Y) = (\beta \cdot \alpha + \gamma(\beta))^2 + \beta^2 \cdot (\Sigma_X - \alpha^2) + \sigma^2_Y(\beta) \Rightarrow \sigma^2_Y(\beta) = \Sigma_Y - (\beta\alpha + \gamma(\beta))^2 - \beta^2(\Sigma_X - \alpha^2)$$

Per ogni β per cui $\sigma^2_Y(\beta) > 0$ e $\gamma(\beta)$ è ben definito, il modello M_β ha distribuzione osservazionale esattamente $(\Sigma_X, \Sigma_{\{XY\}}, \Sigma_Y)$ e produce l'effetto causale:

$$\mathbb{E}[Y \mid \text{do}(X=x)] = \beta \cdot x$$

La condizione $\sigma^2_Y(\beta) > 0$ impone un vincolo su β che è tuttavia soddisfatto in un intorno aperto di $\beta = 0$ per valori tipici dei parametri: calcolando esplicitamente $\sigma^2_Y(0) = \Sigma_Y - (\gamma(0))^2 \cdot \alpha^2 \cdot \dots = \Sigma_Y - (\Sigma_{\{XY\}}/\alpha)^2 \cdot \dots > 0$ per α sufficientemente grande rispetto a $\Sigma_{\{XY\}}$. La funzione $\beta \mapsto \sigma^2_Y(\beta)$ è continua e strettamente positiva su un intervallo aperto, l'intervallo ammissibile è non degenere. $\square$

Il commento alla dimostrazione illumina la struttura del meccanismo. Il bias di confondimento nel modello lineare è il termine di omitted variable bias della tradizione econometrica:

$$b_{\text{OLS}} - \beta = \text{Cov}(X, U) \cdot \text{Cov}(U, Y \mid X) / \text{Var}(X) = \alpha \cdot \gamma(\beta) / \Sigma_X = (\Sigma_{\{XY\}} - \beta \cdot \Sigma_X) / \Sigma_X = \Sigma_{\{XY\}} / \Sigma_X - \beta$$

Il coefficiente OLS è $b_{\text{OLS}} = \Sigma_{\{XY\}} / \Sigma_X$ indipendentemente da β , mentre l'effetto causale β varia: l'intera variazione di β viene assorbita dalla variazione di $\gamma(\beta) = (\Sigma_{\{XY\}} - \beta \cdot \Sigma_X) / \alpha$, mantenendo b_{OLS} costante. Questo è il meccanismo esatto per cui il confondimento rende l'effetto causale non identificabile: la distribuzione osservazionale non può distinguere una stima OLS distorta di un effetto causale grande da una stima OLS meno distorta di un effetto causale più piccolo.

III. Identificabilità condizionale al grafo: la formula di aggiustamento per backdoor

Il teorema di non-identificabilità riguarda il caso in cui non si conosce nemmeno il grafo causale G . Quando invece il grafo è noto — quando si dispone di informazione qualitativa sulla struttura causale, derivata dalla teoria del dominio, da esperimenti precedenti, o da assunzioni biologicamente motivate — la situazione è molto più favorevole: l'identificabilità diventa una questione di topologia del grafo, e il criterio del backdoor fornisce una condizione sufficiente elegante e algoritmicamente verificabile.

I percorsi backdoor da X a Y sono i percorsi che iniziano con un arco diretto *entrante* in X (archi che puntano verso X), attraversano il grafo, e raggiungono Y. Questi percorsi rappresentano le vie attraverso cui il confondimento — la dipendenza tra X e Y non mediata da $X \rightarrow Y$ direttamente — fluisce nel grafo. Bloccare questi percorsi tramite un insieme osservato di variabili elimina il confondimento.

Definizione (criterio del backdoor). Un insieme di nodi $Z \subseteq V \setminus \{X, Y\}$ soddisfa il criterio del backdoor rispetto alla coppia (X, Y) nel grafo G se: (i) nessun nodo in Z è un discendente di X nel grafo G; (ii) Z blocca — nel senso della d-separazione — tutti i percorsi backdoor da X a Y nel grafo G.

Quando tale insieme Z esiste, vale la formula di aggiustamento:

$$P(Y \mid \text{do}(X=x)) = \sum_z P(Y \mid X=x, Z=z) \cdot P(Z=z)$$

Dimostrazione rigorosa della formula di aggiustamento per backdoor.

Sia M un SCM con grafo G in cui Z soddisfa il criterio del backdoor rispetto a (X, Y). Si vuole mostrare che la distribuzione di Y nel modello mutilato $M_{\{X \leftarrow x\}}$ coincide con la media ponderata delle probabilità condizionali stratificate per Z, con pesi dati dalla distribuzione marginale di Z.

Passo 1. Per la legge della probabilità totale nel modello mutilato $M_{\{X \leftarrow x\}}$:

$$P_{\{M_{\{X \leftarrow x\}}\}}(Y=y) = \sum_z P_{\{M_{\{X \leftarrow x\}}\}}(Y=y \mid Z=z) \cdot P_{\{M_{\{X \leftarrow x\}}\}}(Z=z)$$

Passo 2: identificazione di $P_{\{M_{\{X \leftarrow x\}}\}}(Z=z)$ con $P(Z=z)$. Nel modello mutilato, gli archi entranti in X sono rimossi. La condizione (i) del criterio del backdoor — nessun nodo in Z è discendente di X — garantisce che la distribuzione di Z nel modello mutilato coincide con la distribuzione osservazionale: eliminando gli archi entranti in X, si modifica la distribuzione di X e dei suoi discendenti, ma Z non è discendente di X, quindi $P_{\{M_{\{X \leftarrow x\}}\}}(Z) = P_M(Z)$.

Passo 3: identificazione di $P_{\{M_{\{X \leftarrow x\}}\}}(Y=y \mid Z=z)$ con $P(Y=y \mid X=x, Z=z)$. Nel modello mutilato $M_{\{X \leftarrow x\}}$, X è una costante fissata a x. La condizione (ii) del criterio — Z d-separa X da Y lungo tutti i percorsi backdoor nel grafo G — significa che, nel grafo mutilato $G_{\{X\}}$ (con archi entranti in X rimossi), X è una sorgente e i soli percorsi tra X e Y sono quelli causali (uscanti da X). Poiché Z blocca i percorsi backdoor nel grafo originale G, nel grafo mutilato $G_{\{X\}}$ i soli percorsi attivi tra X e Y dato Z sono quelli causali. La correlazione residua tra X (fissata a x) e Y dato $Z=z$ è esattamente quella dovuta al percorso causale diretto, che coincide con la relazione osservata condizionando su $X=x$ e $Z=z$ nella distribuzione osservazionale. Pertanto:

$$P_{\{M_{\{X \leftarrow x\}}\}}(Y=y \mid Z=z) = P_M(Y=y \mid X=x, Z=z)$$

Passo 4. Combinando i passi 1–3:

$$P(Y=y \mid \text{do}(X=x)) = P_{\{M_{\{X \leftarrow x\}}\}}(Y=y) = \sum_z P(Y=y \mid X=x, Z=z) \cdot P(Z=z) \quad \square$$

Interpretazione della formula di aggiustamento. La struttura è quella di una standardizzazione diretta: si calcolano le relazioni condizionali di Y su X stratificate per Z , e si fa la media con i pesi marginali $P(Z=z)$ — non con i pesi condizionali $P(Z=z|X=x)$ che differiscono tra i valori di X quando X e Z sono correlati. È precisamente questa differenza di pesatura che il paradosso di Simpson esemplifica: la distribuzione osservazionale utilizza pesi $P(Z|X=x)$ dipendenti da x , mentre la formula causale adotta pesi marginali $P(Z)$ che non dipendono dall'intervento su X .

Il post-treatment bias. La condizione (i) — nessun nodo di Z discendente di X — ha conseguenze immediate per la pratica econometrica: è scorretto includere come covariate variabili causate dal trattamento. L'inclusione di variabili post-trattamento come controlli nella regressione produce il *conditioning on a descendant bias*, che distorce la stima dell'effetto causale in una direzione che dipende dalla struttura causale della variabile discendente. In uno studio sull'effetto di un programma di formazione (X) sull'occupazione (Y), includere come covariata lo stato di impiego intermedio (Z , causato dal programma) bloccherebbe parte del percorso causale $X \rightarrow Z \rightarrow Y$, producendo una sottostima dell'effetto totale. Non esiste alcun modo di correggere questo errore aumentando il campione o raffinando la stima numerica: è un errore di specificazione del modello causale, irreparabile senza modificare la struttura delle covariate.

IV. Il criterio del front-door e l'identificabilità attraverso mediatori

Il criterio del backdoor gestisce il caso in cui esiste un insieme osservabile di variabili che blocca tutti i percorsi confondenti. Esistono tuttavia configurazioni in cui questo non è possibile — dove il confondimento è generato da variabili latenti non osservabili e non esiste alcun insieme Z osservabile che soddisfi il criterio del backdoor — ma in cui l'effetto causale rimane identificabile attraverso un argomento più sottile, che sfrutta la struttura del percorso causale attraverso mediatori osservabili.

Definizione (criterio del front-door). Un insieme di nodi $M \subseteq V \setminus \{X, Y\}$ soddisfa il criterio del front-door rispetto alla coppia (X, Y) nel grafo G se: (i) M intercetta tutti i percorsi causali diretti da X a Y — ogni percorso causale da X verso Y nel DAG passa attraverso qualche nodo in M ; (ii) non vi sono percorsi backdoor non bloccati da X verso M ; (iii) tutti i percorsi backdoor da M verso Y sono bloccati da X .

Quando tali condizioni sono soddisfatte, vale la formula front-door:

$$P(Y \mid \text{do}(X=x)) = \sum_m P(M=m \mid X=x) \cdot \sum_{x'} P(Y \mid M=m, X=x') \cdot P(X=x')$$

Derivazione della formula front-door. Per la condizione (i), tutti i percorsi causali da X a Y passano per M . Si fattorizza l'effetto di X su Y attraverso gli effetti separati $X \rightarrow M$ e $M \rightarrow Y$:

$$P(Y \mid \text{do}(X=x)) = \sum_m P(Y \mid \text{do}(X=x), \text{do}(M=m)) \cdot P(M=m \mid \text{do}(X=x))$$

Poiché M intercetta tutti i percorsi causali da X a Y , una volta fissato M con un intervento separato l'effetto di X su Y è nullo. Pertanto:

$$P(Y \mid \text{do}(X=x), \text{do}(M=m)) = P(Y \mid \text{do}(M=m))$$

Il termine $P(M=m \mid \text{do}(X=x))$ è identificabile grazie alla condizione (ii): non vi sono percorsi backdoor non bloccati da X verso M , quindi:

$$P(M=m \mid \text{do}(X=x)) = P(M=m \mid X=x)$$

Il termine $P(Y \mid \text{do}(M=m))$ si identifica tramite il criterio del backdoor con X come insieme di aggiustamento, per la condizione (iii):

$$P(Y \mid \text{do}(M=m)) = \sum_{x'} P(Y \mid M=m, X=x') \cdot P(X=x')$$

Combinando i tre passaggi:

$$P(Y \mid \text{do}(X=x)) = \sum_m P(M=m \mid X=x) \cdot \sum_{x'} P(Y \mid M=m, X=x') \cdot P(X=x') \quad \square$$

L'esempio paradigmatico del tabacco e del catrame. L'applicazione più celebre del criterio front-door riguarda la stima dell'effetto causale del fumo (X) sul cancro al polmone (Y) in presenza di un gene latente (U) che predispone sia al fumo sia al cancro. Questo scenario è particolarmente significativo nella storia dell'epidemiologia: Ronald Fisher negli anni Cinquanta contestò l'evidenza epidemiologica del legame fumo-cancro invocando precisamente questo tipo di confondimento genetico. Il catrame nei polmoni (M) è un mediatore osservabile: il fumo deposita catrame, il catrame causa mutazioni cancerogene. Se il catrame soddisfa le condizioni del criterio front-door, allora la formula front-door permette di stimare l'effetto causale del fumo sul cancro controllando per il catrame senza dover osservare U — confutando l'obiezione di Fisher su basi puramente osservazionali. Questo è logicamente sorprendente: la struttura causale del percorso mediato consente di aggirare il confondimento latente che una strategia backdoor non potrebbe neutralizzare.

V. Il do-calculus: sistema completo e teorema di completezza

Il criterio del backdoor e il criterio del front-door sono due strumenti specifici per due classi di grafi. Pearl ha dimostrato che l'intera teoria dell'identificabilità — per qualsiasi grafo, qualsiasi distribuzione osservazionale, qualsiasi quantità causale — può essere catturata da un sistema completo di tre regole di trasformazione: il do-calculus.

Richiamiamo la notazione necessaria. Per due insiemi A e B separati dall'insieme C nel grafo G si scrive $(A \perp_d B \mid C)_G$. La notazione $G\{X\}$ indica il grafo G con tutti gli archi entranti in X rimossi; $G\{X\bar{Z}\}$ indica la rimozione degli archi entranti sia in X sia in Z ; $G_{-}\{Z(W)\}$ indica la rimozione degli archi uscenti da Z verso i nodi dell'insieme W che non sono antenati di Y nel sottografo $G_{-}\{X\}$.

Regola 1 (inserzione e rimozione di osservazioni). Se $(Y \perp_d Z \mid X, W)$ nel grafo $G\{X\}$, allora:

$$P(Y \mid \text{do}(X=x), Z=z, W=w) = P(Y \mid \text{do}(X=x), W=w)$$

Interpretazione. L'osservazione di Z è irrilevante per Y dopo l'intervento su X quando Z e Y sono d -separati nel grafo mutilato. Il condizionamento su informazione irrilevante non altera la distribuzione post-intervento.

Regola 2 (scambio tra osservazione e intervento). Se $(Y \perp_d Z \mid X, W)$ nel grafo $G\{X\bar{Z}\}$, allora:

$$P(Y \mid \text{do}(X=x), \text{do}(Z=z), W=w) = P(Y \mid \text{do}(X=x), Z=z, W=w)$$

Interpretazione. Quando la d -separazione è soddisfatta nel grafo doppiamente mutilato, intervenire su Z produce lo stesso effetto che osservare Z . La condizione è verificata quando Z non ha percorsi backdoor verso Y nel contesto dell'intervento su X , garantendo che osservare Z equivale a un intervento su Z .

Regola 3 (inserzione e rimozione di interventi). Se $(Y \perp_d Z \mid X, W)$ nel grafo $G\{X, Z(W)\}$, dove $Z(W)$ denota la rimozione degli archi uscenti da Z verso $W(Z)$ — l'insieme dei nodi di W che non sono antenati di Y nel sottografo $G_{-}\{X\}$ privato degli archi entranti in Z — allora:

$$P(Y \mid \text{do}(X=x), \text{do}(Z=z), W=w) = P(Y \mid \text{do}(X=x), W=w)$$

Interpretazione. Sotto opportune condizioni topologiche, l'intervento su Z è irrilevante per Y e può essere eliminato dall'espressione. La condizione garantisce che Z non influenzi Y né direttamente né indirettamente attraverso W , nel contesto dell'intervento su X .

Derivazione delle regole. Le tre regole derivano dalla proprietà di Markov del modello mutilato. Il modello mutilato $M_{-}\{X \leftarrow x\}$ ha una distribuzione $P_{-}\{M_{-}\{X \leftarrow x\}\}$ che soddisfa la condizione di Markov rispetto al grafo mutilato $G_{-}\{X\}$. Le regole del do-calculus esprimono le implicazioni di questa condizione di Markov in termini di manipolazioni algebriche sulle espressioni che coinvolgono l'operatore $\text{do}(\cdot)$. La soundness di ogni regola — il fatto che la trasformazione preservi la correttezza dell'uguaglianza — si dimostra verificando che le ipotesi di d -separazione garantiscono le indipendenze statistiche richieste nel modello mutilato corrispondente. Le derivazioni complete sono condotte nell'Appendice A.

Teorema di completezza del do-calculus (Huang e Valtorta 2006; Shpitser e Pearl 2006). Una quantità causale Q è identificabile dalla distribuzione osservazionale $P(V)$ in un ADMG G se e solo se esiste una sequenza finita di applicazioni delle tre regole del do-calculus che trasforma Q in un'espressione contenente soltanto distribuzioni osservazionali $P(V)$ senza operatori $\text{do}(\cdot)$.

Il teorema di completezza ha due conseguenze di prima importanza. La prima, positiva: per ogni quantità causale identificabile, il do-calculus fornisce una procedura algoritmica per calcolarne la formula di identificazione. La seconda, negativa: se il sistema non riesce a trasformare Q in forma osservazionale, allora Q non è identificabile — non esiste formula statistica che la calcoli, indipendentemente dalla sofisticazione dell'analisi. Il do-calculus trasforma l'identificabilità causale da questione aperta ad algoritmo di decisione.

VI. La caratterizzazione della non-identificabilità: c-componenti e hedge

La completezza del do-calculus permette di caratterizzare la non-identificabilità in modo algoritmico. La teoria delle c-componenti di Tian e Pearl, e il concetto di hedge di Shpitser e Pearl, forniscono l'immagine geometrica del perché certi grafi non permettono l'identificazione e certificano la non-identificabilità in modo strutturalmente esplicito.

Definizione (ADMG e archi bidirezionali). Un ADMG (Acyclic Directed Mixed Graph) è un grafo che contiene sia archi diretti ($\rightarrow$) sia archi bidirezionali ($\leftrightarrow$). Gli archi bidirezionali $V_i \leftrightarrow V$ rappresentano l'esistenza di un confounder latente comune che causa sia V_i sia V — sono la proiezione sulla distribuzione osservabile di un percorso $V_i \leftarrow U \rightarrow V$ con U latente.

Definizione (c-componente). In un ADMG G , due nodi V_i e V appartengono alla stessa c-componente se sono connessi da un percorso che consiste soltanto di archi bidirezionali. Le c-componenti partizionano i nodi del grafo in sottoinsiemi massimali connessi tramite archi bidirezionali. Intuitivamente, una c-componente raggruppa le variabili che condividono, direttamente o transitivamente, variabili latenti comuni.

Definizione (hedge). Una coppia di insiemi di nodi (F, F') forma una hedge per la coppia (X, Y) in un ADMG G se: (i) $X \in F$ e $X \notin F'$; (ii) $F' \subseteq F$; (iii) sia F sia F' sono c-componenti del sottografo $G[F]$ — il sottografo di G indotto dai nodi in F ; (iv) $F' = \text{An}(Y)_{-}\{G[V \setminus X]\}$ — l'insieme degli antenati di Y nel sottografo di G ottenuto rimuovendo X e tutti i suoi archi uscenti.

L'intuizione geometrica della hedge è la seguente: gli insiemi F e F' formano una struttura a siepe intorno all'effetto causale di X su Y . F' è l'insieme di variabili antenate di Y nel grafo privato di X — le variabili causalmente rilevanti per Y che non sono direttamente influenzate da X nell'intervento. F è un insieme più grande che include X stesso. Il fatto che F e F' appartengano alla stessa c-componente di $G[F]$ significa che X e la struttura causale intorno ad essa sono connesse tramite cause latenti condivise in modo tale che l'intervento su X non può essere separato dall'influenza di quelle latenti sulla distribuzione di Y .

Teorema (Shpitser e Pearl, 2006 — caratterizzazione della non-identificabilità). $P(Y \mid \text{do}(X=x))$ è non-identificabile in G se e soltanto se G contiene una hedge per (X, Y) .

Questo teorema fornisce un criterio geometrico-combinatorio: basta trovare la struttura caratteristica della hedge nel grafo per concludere che nessun aggiustamento statistico può recuperare $P(Y \mid \text{do}(X))$ dalla distribuzione osservazionale. La presenza di una hedge è verificabile in tempo polinomiale nel numero di nodi, fornendo un test efficiente di non-identificabilità.

L'algoritmo ID di Shpitser e Pearl combina il teorema di completezza del do-calculus con il criterio della hedge: dato un ADMG G e una quantità causale Q , l'algoritmo decide in tempo polinomiale se Q è identificabile; se lo è, restituisce la formula di identificazione; se non lo è, individua la hedge che certifica la non-identificabilità. L'algoritmo ID è la soluzione completa al problema dell'identificabilità causale per DAG e ADMG.

VII. La transizione al livello-3: non-identificabilità dei controfattuali

La gerarchia causale ha una terza frontiera, più profonda della precedente, che separa le distribuzioni interventistiche (livello-2) dalle probabilità controfattuali (livello-3). Il risultato fondamentale qui è che le distribuzioni post-intervento non sono sufficienti per determinare i controfattuali: anche se tutte le distribuzioni $\{P(V \mid \text{do}(X_i=x_i))\}$ per tutti i possibili interventi sono note esattamente, le probabilità controfattuali $P(Y_{\leftarrow x} = y \mid X=x', Y=y')$ rimangono non-identificabili in generale.

Definizione (controfattuale). Nel modello causale strutturale M , il controfattuale $Y_{\leftarrow x}(u)$ denota il valore che Y assumerebbe per l'individuo u se X fosse fissato a x per intervento, a partire dalla situazione in cui u è l'individuo reale con le sue variabili esogene fissate al valore u . Formalmente:

$$Y_{\leftarrow x}(u) = f_Y(\text{PA}(Y) \text{ nel modello mutilato } M_{\leftarrow x} \text{ con } U=u)$$

La probabilità controfattuale $P(Y_{\leftarrow x}=y \mid X=x', Y=y')$ è la probabilità che il valore controfattuale di Y sia y per individui che nella realtà hanno $X=x'$ e $Y=y'$. Il calcolo richiede la procedura di abduzione-azione-predizione di Pearl:

Abduzione: aggiornare la distribuzione di U sulla base dell'evidenza $(X=x', Y=y')$, ottenendo $P(U \mid X=x', Y=y')$ tramite il teorema di Bayes.

Azione: costruire il modello mutilato $M_{\leftarrow x}$.

Predizione: calcolare la distribuzione di Y nel modello mutilato $M_{\leftarrow x}$ usando la distribuzione di U aggiornata per l'evidenza:

$$P(Y_{\leftarrow x}=y \mid X=x', Y=y') = \int \mathbf{1}[f_Y(\text{PA}(Y), u) = y \text{ nel modello } M_{\leftarrow x}] \cdot P(U=u \mid X=x', Y=y') d\mu(u)$$

Il passo cruciale di abduzione richiede la conoscenza della distribuzione congiunta $P(U)$ e delle equazioni strutturali in forma funzionale completa — non solo della distribuzione marginale di V o delle distribuzioni interventistiche.

Teorema (non-identificabilità dei controfattuali da distribuzioni interventistiche).

Esistono due SCM M_1 e M_2 tali che $P_{\leftarrow M_1}(V \mid \text{do}(S=s)) = P_{\leftarrow M_2}(V \mid \text{do}(S=s))$ per ogni sottoinsieme $S \subseteq V$ e ogni valore s , ma $P_{\leftarrow M_1}(Y_{\leftarrow x}=y \mid X=x', Y=y') \neq P_{\leftarrow M_2}(Y_{\leftarrow x}=y \mid X=x', Y=y')$ per almeno una combinazione (y, x, x', y') .

Prova per costruzione. Si considerano due modelli bivariati con X binaria e Y continua:

M_1 : $X := U_X, Y := X + U_Y$, con $U_X \sim \text{Bernoulli}(1/2)$ e $U_Y \sim N(0,1)$ indipendenti.

M_2 : $X := U_X, Y := X + h(X, U_Y)$, con la funzione h scelta in modo da soddisfare: $h(0, u_Y) = u_Y$ per ogni u_Y (funzione identità quando $X=0$) e $h(1, u_Y) = T(u_Y)$ per una trasformazione biettiva T tale che $T(U_Y) \sim N(0,1)$ ma $T \neq \text{identità}$.

Le distribuzioni marginali di (X, Y) e le distribuzioni interventistiche $P(Y \mid \text{do}(X=x))$ coincidono tra M_1 e M_2 per la scelta appropriata di T (che preserva la distribuzione marginale di $Y \mid X=1$ tramite la biettività di T e la gaussianità di $T(U \mid Y)$). Tuttavia le probabilità controfattuali differiscono: per un individuo osservato con $(X=0, Y=y_0)$, il valore controfattuale $Y_{\{X \leftarrow 1\}}$ in M_1 è $y_0 + 1$ (poiché $U \mid Y = y_0$ nel modello mutilato dà $Y = 1 + y_0$), mentre in M_2 è $T(y_0) + 1$. Poiché $T \neq \text{identità}$, si ha $T(y_0) \neq y_0$ per quasi ogni y_0 , e le probabilità controfattuali differiscono quasi certamente. $\square$

La ragione strutturale della non-identificabilità dei controfattuali è la seguente: i controfattuali dipendono dalla *struttura di dipendenza congiunta* tra le variabili esogene U condizionate all'evidenza osservata. Questa struttura congiunta non è determinata dalle distribuzioni marginali interventistiche — che sono distribuzioni di V per valori fissi di un sottoinsieme di V , non distribuzioni di U . Le distribuzioni interventistiche forniscono informazione sulle medie e varianze di U attraverso le funzioni strutturali, ma non sull'intera struttura di dipendenza di U condizionata all'evidenza individuale — la quantità che determina i controfattuali.

VI. La caratterizzazione geometrica della non-identificabilità: c-componenti e hedge

La completezza del do-calculus garantisce che ogni quantità causale identificabile possa essere trasformata in forma osservazionale tramite le tre regole. Ma questa caratterizzazione è algoritmica, non strutturale: non offre un'immagine geometrica del perché certi grafi non permettono l'identificazione. La teoria delle c-componenti di Tian e Pearl (2002), e il concetto di hedge di Shpitser e Pearl (2006), forniscono questa immagine strutturale e trasformano la non-identificabilità da risposta algoritmicamente prodotta a proprietà geometricamente visibile nel grafo.

VI.1 ADMG e archi bidirezionali: la proiezione delle variabili latenti

Un ADMG (Acyclic Directed Mixed Graph) è un grafo che contiene sia archi diretti $V_i \rightarrow V_j$ (relazioni causali osservabili) sia archi bidirezionali $V_i \leftrightarrow V_j$ (cause latenti comuni non osservabili). Gli archi bidirezionali sono la proiezione sulla distribuzione osservabile di un percorso $V_i \leftarrow U \rightarrow V_j$ con U latente non misurata.

La costruzione formale è la seguente. Dato un DAG completo G^* su variabili osservate $V = \{V_1, \dots, V_n\}$ e variabili latenti $L = \{U_1, \dots, U_m\}$, l'ADMG di proiezione G su V è ottenuto:

- (a) Mantenendo ogni arco diretto $V_i \rightarrow V_j$ che esiste in G^* tra variabili osservate;
- (b) Per ogni coppia (V_i, V_j) di variabili osservate che hanno un antenato latente comune U - cioè $V_i \leftarrow \dots \leftarrow U \rightarrow \dots \rightarrow V_j$ in G^* con $U \in L$ - inserendo un arco bidirezionale $V_i \leftrightarrow V_j$ in G .

La distribuzione osservazionale $P(V)$ generata dal DAG completo G^* coincide con quella generata dall'ADMG G : il modello marginalizzato sulle latenti ha esattamente le stesse implicazioni di m-separazione (la generalizzazione della d-separazione ai grafi misti) del modello completo, purché G sia massimale - ogni coppia non adiacente in G è m-separata da qualche insieme.

VI.2 C-componenti: definizione e proprietà fondamentali

Definizione (c-componente). In un ADMG G , due nodi V_i e V_j appartengono alla stessa c-componente se sono connessi da un percorso che consiste soltanto di archi bidirezionali. Formalmente, la relazione $\sim$ definita da $V_i \sim V_j$ sse $\exists$ percorso $V_i \leftrightarrow \dots \leftrightarrow V_j$ in G è una relazione di equivalenza, e le c-componenti sono le classi di equivalenza di $\sim$.

La struttura algebrica fondamentale delle c-componenti è la seguente decomposizione di Tian e Pearl (2002):

Teorema (Tian e Pearl, 2002 - decomposizione in c-fattori). Sia G un ADMG con insieme di nodi V e c-componenti $S_1, \dots, S_k$. Allora la distribuzione osservazionale $P(V)$ si decompone come:

$$P(V) = \prod_{j=1}^k Q[S_j]$$

dove $Q[S_j]$ è il c-fattore della componente S_j , definito come:

$$Q[S_j] = \sum P(V_k \mid V_{\setminus \{k-1\}}, \dots, V_{-1}) \text{ for } V_k \text{ in } S_j$$

con π un ordinamento topologico delle variabili endogene, e la sommatoria è su tutte le variabili non in S_j . Ogni c-fattore $Q[S_j]$ è univocamente identificabile dalla distribuzione osservazionale $P(V)$ - questa è la chiave dell'identificabilità parziale.

Dimostrazione della decomposizione. La dimostrazione procede per induzione sulla struttura del grafo.

Passo base: Se G non ha archi bidirezionali, ogni nodo è la propria c-componente. La decomposizione diventa $P(V) = \prod_i P(V_i \mid \text{PA}(V_i))$, la fattorizzazione di Markov standard del DAG. Ogni fattore è identificabile dalla distribuzione condizionale osservata.

Passo induttivo: Supponiamo che G abbia almeno un arco bidirezionale. Sia T la c-componente che contiene il nodo con rango massimale nell'ordinamento topologico π . Per la struttura dell'ADMG, gli archi entranti in T dall'esterno sono tutti archi diretti (non bidirezionali), e quindi le variabili in $V \setminus T$ non condividono variabili latenti con quelle in T attraverso percorsi che attraversino il bordo di T . Per il principio di modularità, si può fattorizzare $Q[T]$ separatamente.

La prova formale completa usa la teoria delle marginali di distribuzione causale e il criterio di m-separazione; rimandiamo a Tian e Pearl (2002, Teorema 3) per i dettagli tecnici. $\square$

VI.3 La nozione di hedge: definizione formale e intuizione geometrica

Definizione (hedge). Una coppia di insiemi di nodi (F, F') forma una hedge per la coppia (X, Y) in un ADMG G se:

- (i) $X \in F$ e $X \notin F'$;
- (ii) $F' \subseteq F$;
- (iii) Sia F sia F' sono c-componenti del sottografo $G[F]$ - il sottografo di G indotto dai nodi in F ;
- (iv) $F' = \text{An}(Y)_{\setminus \{G[V \setminus X]\}}$ - l'insieme degli antenati di Y nel sottografo di G ottenuto rimuovendo X e tutti i suoi archi uscenti, intersecato con F .

Intuizione geometrica. Una hedge struttura lo spazio causale attorno all'effetto di X su Y come una "doppia siepe". L'insieme F' rappresenta la zona di influenza di Y che è causalmente accessibile senza passare per X : sono le variabili che Y "vede" indipendentemente dall'intervento su X . L'insieme $F \supseteq F'$ è più grande e include X stesso, rappresentando l'intera zona influenzata da X insieme alle sue "radici" latenti.

L'appartenenza sia di F che di F' alla stessa c-componente del sottografo $G[F]$ è la condizione critica: essa afferma che X e la struttura F' condividono variabili latenti comuni attraverso la struttura di confondimento in F . Questo crea un "cortocircuito" causale: l'intervento $\text{do}(X=x)$ taglia i percorsi da X verso Y , ma la latente comune U con $U \rightarrow X$ e $U \rightarrow F'$ re-introduce la dipendenza attraverso il confondimento.

VI.4 Esempi di hedge: strutture tipiche di non-identificabilità

Esempio 1: Bow-arc (la struttura più semplice). Consideriamo il grafo G con due nodi X e Y , un arco $X \rightarrow Y$ e un arco bidirezionale $X \leftrightarrow Y$. Gli archi bidirezionali codificano la presenza di una variabile latente U con $U \rightarrow X$ e $U \rightarrow Y$.

Poniamo $F = \{X, Y\}$ e $F' = \{Y\}$. Verifichiamo le condizioni:

- (i) $X \in \{X, Y\}$ e $X \notin \{Y\}$: verificato.
- (ii) $F' = \{Y\} \subseteq \{X, Y\} = F$: verificato.
- (iii) Nel sottografo $G[\{X, Y\}]$, gli archi bidirezionali presenti sono $X \leftrightarrow Y$: $F = \{X, Y\}$ è una c-componente (connessa tramite $\leftrightarrow$) e $F' = \{Y\}$ è una c-componente di $G[\{X, Y\}]$ (il nodo singleton Y è sempre una c-componente di qualsiasi sottografo).
- (iv) $\text{An}(Y)_{\setminus \{G[V \setminus X]\}} = \{Y\}$ (senza X , il solo antenato di Y nel grafo residuo è Y stesso, poiché l'unico percorso diretto verso Y era $X \rightarrow Y$): verificato.

Dunque $(\{X,Y\}, \{Y\})$ è una hedge per (X, Y) in G . Per il teorema di Shpitser e Pearl, $P(Y \mid \text{do}(X=x))$ non è identificabile in questo grafo - un risultato intuitivo, poiché U confonde la relazione X - Y senza variabili di aggiustamento disponibili.

Esempio 2: Grafo a tre nodi con hedge non banale. Consideriamo G con nodi $\{X, M, Y\}$, archi $X \rightarrow M$, $M \rightarrow Y$, e archi bidirezionali $X \leftrightarrow Y$ e $M \leftrightarrow Y$. Si chiede se $P(Y \mid \text{do}(X=x))$ sia identificabile.

Tentiamo di costruire una hedge. Poniamo $F = \{X, M, Y\}$. Nel sottografo $G[\{X,M,Y\}]$, gli archi bidirezionali sono $X \leftrightarrow Y$ e $M \leftrightarrow Y$. Le c-componenti di $G[\{X,M,Y\}]$ sono: $\{X, Y, M\}$ (tutti connessi da $\leftrightarrow$ attraverso il nodo Y). Dunque F è una c-componente.

Poniamo $F' = \text{An}(Y) \setminus \{G[V \setminus X]\} \cap F$. Rimuovendo X , il sottografo $G[\{M,Y\}]$ ha $M \rightarrow Y$ e $M \leftrightarrow Y$. Gli antenati di Y senza X sono $\{M, Y\}$. Dunque $F' = \{M, Y\}$. Verifichiamo la condizione (iii): $F' = \{M,Y\}$ deve essere c-componente di $G[F] = G[\{X,M,Y\}]$. In $G[\{X,M,Y\}]$, M e Y sono connessi da $\leftrightarrow$, quindi $\{M,Y\}$ è effettivamente una c-componente.

$(\{X,M,Y\}, \{M,Y\})$ è quindi una hedge per (X,Y) : $P(Y \mid \text{do}(X=x))$ è non-identificabile in questo grafo. In particolare, il criterio del front-door non si applica ($M \leftrightarrow Y$ viola la condizione di no-backdoor da M verso Y), e il do-calculus non riesce a trasformare $P(Y \mid \text{do}(X=x))$.

VI.5 Il teorema di caratterizzazione della non-identificabilità

Teorema (Shpitser e Pearl, 2006). Sia G un ADMG e (X, Y) una coppia di sottoinsiemi di nodi. Allora $P(Y \mid \text{do}(X=x))$ è non-identificabile dalla distribuzione osservazionale $P(V)$ se e soltanto se G contiene una hedge per (X, Y) .

Schema della dimostrazione. La dimostrazione procede in due direzioni.

Direzione ($\Leftarrow$): la presenza di una hedge implica la non-identificabilità. Supponiamo che (F, F') sia una hedge per (X, Y) in G . Si costruisce una famiglia di SCM parametrizzata da un parametro reale λ , tutti compatibili con G e con la stessa distribuzione osservazionale $P(V)$, ma con distribuzioni interventistiche $P(Y \mid \text{do}(X=x))$ diverse al variare di λ .

La costruzione sfrutta la struttura della hedge: poiché F e F' appartengono alla stessa c-componente di $G[F]$, esiste una variabile latente U con influenza su X (in F) e su F' . Si parametrizza l'intensità dell'influenza di U su Y attraverso F' con il parametro λ , e si verifica che la distribuzione osservazionale è invariante in λ mentre $P(Y \mid \text{do}(X=x)) = f(\lambda)$ varia. Questo segue dal teorema di incommensurabilità della Sezione VI della Parte Prima, applicato alla struttura locale della hedge.

Direzione ($\Rightarrow$): l'assenza di hedge implica l'identificabilità. Si dimostra per induzione sulla struttura dell'ADMG che se G non contiene hedge, allora l'algoritmo ID di Shpitser e Pearl termina producendo una formula di identificazione. L'algoritmo ID è una serie di passi che decompone $P(Y \mid \text{do}(X=x))$ in sotto-problemi più semplici usando la struttura delle c-componenti: a ogni passo, o si applica una regola del do-calculus che semplifica

l'espressione, o si costruisce la decomposizione in c-fattori che riduce il problema a sotto-componenti identificabili. Se nessun passo fallisce, il prodotto dell'algoritmo è la formula di identificazione. Se l'algoritmo fallisce, il punto di fallimento corrisponde esattamente a una hedge nel grafo - la struttura che il fallimento "certifica". $\square$

Corollario algoritmico (algoritmo ID). Dato un ADMG G e una coppia (X, Y) , l'algoritmo ID:

1. Verifica se G contiene una hedge per (X, Y) mediante una ricerca in ampiezza sul grafo dei c-fattori.
2. Se non vi è hedge: produce la formula di identificazione di $P(Y \mid \text{do}(X=x))$ come espressione in termini di $P(V)$ in tempo $O(|V|^3)$.
3. Se vi è hedge: individua la coppia (F, F') e certifica la non-identificabilità.

La complessità $O(|V|^3)$ segue dal fatto che la verifica delle condizioni della hedge richiede al più $|V|^2$ verifiche di c-componenti, ciascuna in tempo $O(|V|)$.

VI.6 La struttura analitica dei c-fattori e l'algoritmo di identificazione

L'algoritmo ID opera sulla struttura dei c-fattori in modo ricorsivo. Per ogni c-componente S di G , il c-fattore $Q[S]$ è dato dalla formula:

$$Q[S] = \sum_{\{V_{-1}, \dots, V_{-1}\}} \prod_{\{V_k \in S\}} P(V_k \mid V_{-1}, \dots, V_{-1})$$

dove l'ordinamento è quello topologico π . Quando si applica un intervento $\text{do}(X=x)$, il c-fattore delle componenti che non contengono X rimane invariato; quello della componente contenente X viene modificato dalla rimozione degli archi entranti in X .

Il passo fondamentale dell'algoritmo ID è la seguente riduzione: data una c-componente S che contiene sia variabili di X sia variabili di Y , se S non contiene una hedge interna, allora $Q[S \mid \text{do}(X \cap S = x)]$ è identificabile dal c-fattore $Q[S]$ osservazionale tramite la formula:

$$Q[S \mid \text{do}(X_S = x_S)] = \sum_{\{S \setminus Y\}} Q[S] / Q[X_S]$$

dove $Q[X_S]$ è il c-fattore marginale di X_S . Questa formula è il nucleo computazionale dell'identificazione tramite c-componenti: riduce il problema interventistico a operazioni algebriche su distribuzioni osservazionali.

VII. La transizione al livello-3: non-identificabilità dei controfattuali con trattazione formale completa

La barriera tra livello-2 e livello-3 ha una struttura fondamentale diversa da quella tra livello-1 e livello-2. La barriera 1->2 è una questione di topologia del grafo: alcuni grafi permettono l'identificazione, altri no, e il criterio è geometrico (la presenza di hedge). La barriera 2->3 è una questione di forma funzionale delle equazioni strutturali e di struttura di dipendenza delle variabili esogene: anche se tutte le distribuzioni interventistiche $P(V \mid do(X=x))$ per tutti i possibili interventi sono note esattamente, i controfattuali rimangono non-identificabili in generale.

VII.1 I controfattuali e la procedura di abduzione-azione-predizione

Definizione formale del controfattuale. Nel modello causale strutturale $M = \langle U, V, \mathcal{F}, P(U) \rangle$, il controfattuale $Y_{\{X<-x\}}(u)$ denota il valore che Y assumerebbe per l'individuo caratterizzato dal vettore di variabili esogene $U = u$, se X fosse fissato a x per intervento. Formalmente:

$$Y_{\{X<-x\}}(u) = f_Y(PA(Y) \text{ nel modello mutilato } M_{\{X<-x\}} \text{ con } U=u)$$

dove $PA(Y)$ sono i genitori causali di Y nel DAG G , e nel modello $M_{\{X<-x\}}$ l'equazione strutturale di X è sostituita dalla costante $X := x$.

La procedura di abduzione-azione-predizione. Il calcolo della probabilità controfattuale $P(Y_{\{X<-x\}} = y \mid X=x', Y=y')$ - la probabilità che Y avrebbe assunto valore y per individui che osserviamo avere $X=x'$ e $Y=y'$ - richiede tre passi:

Passo 1 (Abduzione): aggiornare la distribuzione di U sulla base dell'evidenza osservata ($X=x', Y=y'$) mediante il teorema di Bayes:

$$P(U=u \mid X=x', Y=y') = P(X=x', Y=y' \mid U=u) \cdot P(U=u) / P(X=x', Y=y')$$

Questo passo "personalizza" la distribuzione delle variabili esogene: dopo aver osservato $X=x'$ e $Y=y'$, si restringe la famiglia di individui a quelli compatibili con queste osservazioni.

Passo 2 (Azione): costruire il modello mutilato $M_{\{X<-x\}}$, sostituendo l'equazione strutturale di X con la costante $X := x$.

Passo 3 (Predizione): calcolare la distribuzione di Y nel modello $M_{\{X<-x\}}$ con la distribuzione di U aggiornata per l'evidenza:

$$P(Y_{\{X<-x\}}=y \mid X=x', Y=y') = \int 1[f_Y(PA(Y), u) = y \mid M_{\{X<-x\}}] \cdot P(U=u \mid X=x', Y=y') d\mu(u)$$

Il passo critico è l'abduzione: richiede la conoscenza della distribuzione congiunta $P(U)$ e delle equazioni strutturali $\mathcal{F}$ nella loro forma funzionale completa - non solo della distribuzione marginale $P(V)$ o delle distribuzioni interventistiche $P(V \mid do(\cdot))$.

VII.2 Il teorema fondamentale di non-identificabilità dei controfattuali

Teorema (non-identificabilità dei controfattuali da distribuzioni interventistiche).

Esistono due SCM M_1 e M_2 tali che:

(a) $P_{M_1}(V \mid \text{do}(S=s)) = P_{M_2}(V \mid \text{do}(S=s))$ per ogni sottoinsieme $S \subseteq V$ e ogni valore $s \in \mathcal{X}_S$ (identiche distribuzioni interventistiche).

(b) $P_{M_1}(Y_{X \leftarrow x} = y \mid X=x', Y=y') \neq P_{M_2}(Y_{X \leftarrow x} = y \mid X=x', Y=y')$ per almeno una combinazione (y, x, x', y') (controfattuali diversi).

Costruzione formale della prova. Costruiamo i due modelli esplicitamente nel caso più semplice: $X \rightarrow Y$ senza confondimento latente, con X binaria e Y continua.

Modello M_1 : $X := U_X$ con $U_X \sim \text{Bernoulli}(1/2)$; $Y := X + U_Y$ con $U_Y \sim N(0,1)$ indipendente da U_X .

Modello M_2 : $X := U_X$ con $U_X \sim \text{Bernoulli}(1/2)$; $Y := X + h(X, U_Y)$ dove h è definita come:

$$h(0, u) = u \quad [\text{quando } X=0, Y \text{ ha lo stesso valore che in } M_1]$$

$$h(1, u) = T(u) \quad [\text{quando } X=1, Y \text{ è una trasformazione biettiva di } U_Y]$$

dove $T: \mathbb{R} \rightarrow \mathbb{R}$ è una biiezione misurabile tale che $T(U_Y) \sim N(0,1)$ (T preserva la distribuzione marginale di $Y \mid X=1$) ma $T \neq \text{identità}$. Un esempio esplicito: $T(u) = -u$ (riflessione) soddisfa queste condizioni per $U_Y \sim N(0,1)$ poiché $-U_Y \sim N(0,1)$.

Verifica dell'identità osservazionale. Le distribuzioni marginali $P_{M_1}(X,Y)$ e $P_{M_2}(X,Y)$:

Per $X=0$: $Y \mid X=0 \sim N(0,1)$ in entrambi i modelli (identico).

Per $X=1$: in M_1 , $Y \mid X=1 = 1 + U_Y \sim N(1,1)$; in M_2 , $Y \mid X=1 = 1 + T(U_Y) \sim N(1,1)$ per la scelta di T . Identico.

Verifica dell'identità delle distribuzioni interventistiche. $P_{M_1}(Y \mid \text{do}(X=x))$ e $P_{M_2}(Y \mid \text{do}(X=x))$:

Per $\text{do}(X=0)$: $Y = 0 + h(0, U_Y) = U_Y \sim N(0,1)$ in M_2 ; $Y = 0 + U_Y \sim N(0,1)$ in M_1 . Identico.

Per $\text{do}(X=1)$: $Y = 1 + h(1, U_Y) = 1 + T(U_Y) \sim N(1,1)$ in M_2 ; $Y = 1 + U_Y \sim N(1,1)$ in M_1 per la scelta di T . Identico.

Quindi M_1 e M_2 hanno identiche distribuzioni osservazionali e interventistiche.

Calcolo del controfattuale in M_1 e M_2 . Consideriamo un individuo osservato con $X=0$ e $Y=y_0$.

In M_1 : il passo di abduzione dà $U_Y = y_0$ (poiché $Y = 0 + U_Y$ implica $U_Y = y_0$ quando $X=0$). Il controfattuale $Y_{\{X<-1\}}(u_0) = 1 + U_Y = 1 + y_0$.

In M_2 : il passo di abduzione dà $h(0, U_Y) = y_0$, cioè $U_Y = y_0$ (poiché $h(0,u) = u$). Il controfattuale $Y_{\{X<-1\}}(u_0) = 1 + h(1, U_Y) = 1 + T(y_0)$.

Poiché $T \neq$ identità, si ha $T(y_0) \neq y_0$ per quasi ogni y_0 (eccetto al più un insieme di misura zero). Dunque:

$$P_{\{M_1\}}(Y_{\{X<-1\}} = 1+y_0 \mid X=0, Y=y_0) = 1 \neq P_{\{M_2\}}(Y_{\{X<-1\}} = 1+T(y_0) \mid X=0, Y=y_0) = 1$$

I due modelli producono controfattuali diversi per quasi ogni evidenza ($X=0, Y=y_0$), nonostante distribuzioni osservazionali e interventistiche identiche. $\square$

VII.3 Estensione al caso con confondimento latente

La costruzione precedente non usa confondimento latente. Nel caso con confondimento, la non-identificabilità è ancora più radicale perché la distribuzione congiunta di U condizionata all'evidenza dipende dalla struttura del confondimento in modi non accessibili dalle distribuzioni interventistiche marginali.

Costruzione con confondimento $X \leftarrow U \rightarrow Y$ e $X \rightarrow Y$. Le equazioni strutturali sono:

$$U \sim \text{Bernoulli}(1/2) \text{ [latente]}$$

$$X := U \oplus \varepsilon_X \text{ con } \varepsilon_X \sim \text{Bernoulli}(1/2) \text{ indipendente}$$

$$Y := g(X, U, \varepsilon_Y) \text{ con } \varepsilon_Y \sim \text{Bernoulli}(1/2) \text{ indipendente}$$

Consideriamo due scelte di g che producono identiche distribuzioni $P(X,Y)$ e $P(Y \mid \text{do}(X=x))$ ma diversi controfattuali:

Modello M_1 con $g_1(x, u, \varepsilon) = x \oplus u \oplus \varepsilon$. Calcoliamo:

$$P_{\{M_1\}}(Y=1 \mid X=0) = P(0 \oplus U \oplus \varepsilon=1) = P(U \oplus \varepsilon=1) = 1/2. P_{\{M_1\}}(Y=1 \mid X=1) = P(1 \oplus U \oplus \varepsilon=1) = P(U \oplus \varepsilon=0) = 1/2.$$

$$P_{\{M_1\}}(Y=1 \mid \text{do}(X=0)) = P(0 \oplus U \oplus \varepsilon=1) = 1/2. P_{\{M_1\}}(Y=1 \mid \text{do}(X=1)) = P(1 \oplus U \oplus \varepsilon=1) = 1/2.$$

Modello M_2 con $g_2(x, u, \varepsilon) = (x \vee u) \wedge \varepsilon$ (dopo opportuna ricalibrazione di $p = P(\varepsilon=1)$ per uguagliare le distribuzioni). La stessa distribuzione osservazionale e interventistiche è raggiungibile con $p = 1/2$.

Il punto critico è che per un individuo osservato con $(X=0, Y=0)$, il valore del confondimento latente U ha distribuzione diversa nei due modelli (per via della diversa funzione g), e il controfattuale $Y_{\{X<-1\}}$ dipende dalla struttura di dipendenza (U, ε_Y) condizionata all'evidenza - non dalla distribuzione marginale di $Y \mid \text{do}(X=1)$.

VII.4 La ragione strutturale della non-identificabilità: dipendenza congiunta delle esogene

Teorema (strutturale). L'insieme completo delle distribuzioni interventistiche $\{P(V \mid \text{do}(X=x)) : X \subseteq V, x\}$ identifica le distribuzioni marginali di ciascuna variabile esogena U_i e le equazioni strutturali f_i nelle loro statistiche marginali, ma non identifica la struttura di dipendenza congiunta (U_i, U_j) condizionata all'evidenza osservata.

Dimostrazione. Le distribuzioni interventistiche $P(V \mid \text{do}(X_S=x_S))$ per $S \subseteq V$ sono funzionali delle funzioni strutturali f_i e delle distribuzioni marginali $P(U_i)$. Specificamente, ogni $P(V_i \mid \text{do}(X_S=x_S))$ dipende da:

- (a) La funzione $f_i(\text{PA}(V_i), U_i)$ valutata nel modello mutilato (dove gli archi in $\text{PA}(V_i) \cap S$ sono rimossi);
- (b) La distribuzione marginale $P(U_i)$ per ogni variabile esogena U_i presente nella catena causale da X_S a V_i .

La dipendenza (a) identifica f_i tramite le marginali condizionali. La dipendenza (b) identifica $P(U_i)$ singolarmente. Ma le distribuzioni interventistiche $P(V \mid \text{do}(X_S=x_S))$ non contengono alcuna informazione sulla dipendenza congiunta (U_i, U_j) per $i \neq j$ - perché ogni distribuzione interventistica è calcolata nel modello mutilato dove le equazioni strutturali di X_S sono sostituite da costanti, rendendo la distribuzione di U_i e U_j condizionale all'evidenza $X_S=x_S$ indipendente per costruzione.

Il passo di abduzione richiede $P(U \mid X=x', Y=y') = P(U_1, \dots, U_n \mid X=x', Y=y')$: la distribuzione congiunta aggiornata di tutte le esogene. Questa dipendenza congiunta non è determinata dagli esperimenti marginali. $\square$

Implicazione. La frontiera 2->3 della gerarchia causale è invalicabile in linea di principio da qualsiasi quantità di esperimenti interventistici marginali, anche infiniti. Per rispondere a domande controfattuali è necessaria la conoscenza della forma funzionale completa delle equazioni strutturali $\mathcal{F}$ e della struttura di dipendenza congiunta $P(U)$ delle variabili esogene - quantità non accessibili dall'esterno del sistema attraverso osservazioni o interventi marginali.

VII.5 Il teorema di identificabilità parziale dei controfattuali: la classe monotonicità

Non tutti i controfattuali sono non-identificabili. Esiste una classe importante di modelli in cui i controfattuali sono identificabili: i modelli con assunzione di risposta monotonicità.

Definizione (monotonicità). Un modello causale strutturale con X binaria soddisfa la monotonicità se per ogni individuo u : $Y_{\{X<-1\}}(u) \geq Y_{\{X<-0\}}(u)$. In altre parole, l'intervento $X=1$ non può mai peggiorare Y rispetto all'intervento $X=0$.

Teorema (identificabilità dei controfattuali sotto monotonicità). Sotto l'assunzione di monotonicità per una coppia (X, Y) binarie, la probabilità controfattuale $P(Y_{X=1} | X=0, Y=0)$ - la probabilità che Y fosse stato 1 per individui che osserviamo avere X=0 e Y=0 - è identificabile dalla distribuzione osservazionale come:

$$P(Y_{X=1} | X=0, Y=0) = [P(Y=1 | do(X=1)) - P(Y=1 | do(X=0))] / P(Y=0 | X=0)$$

La condizione di monotonicità è la chiave: essa implica che non vi sono individui per cui X=1 causa Y=0 quando X=0 causa Y=1 (gli "always-defiers"). L'identificabilità segue dalla decomposizione della distribuzione interventistica in tipi di risposta.

Limitazione. La monotonicità è un'assunzione di forma funzionale sulla struttura degli errori del modello. Essa non è verificabile dai soli dati osservazionali o interventistici marginali - richiede giudizio teorico sul dominio. La sua violazione annulla immediatamente l'identificabilità dei controfattuali, riportando al risultato di impossibilità generale.

VII.6 Sintesi della frontiera di livello-2->3

Riepilogo strutturale. La non-identificabilità dei controfattuali riflette tre strati distinti di impossibilità che si combinano:

Strato 1 - Forma funzionale. Le equazioni strutturali f_i possono essere equivalenti a livello di distribuzioni marginali osservazionali e interventistiche pur differendo nella struttura funzionale interna. Modelli con diverse f_i "osservazionalmente equivalenti" producono diversi valori controfattuali per gli stessi individui.

Strato 2 - Dipendenza congiunta delle esogene. Due modelli con identiche f_i ma diversa struttura di dipendenza (U_i, U_j) - cioè diverse correlazioni tra gli errori strutturali - producono le stesse distribuzioni interventistiche marginali ma diversi controfattuali attraverso il meccanismo di abduzione.

Strato 3 - Identificazione degli individui. Anche se i livelli precedenti fossero superabili, il passo di abduzione richiede di "riconoscere" l'individuo tramite le variabili esogene u - un'operazione che presuppone la conoscenza della struttura $P(U)$ che le distribuzioni interventistiche non rivelano.

La frontiera 2->3 è dunque invalicabile non per un unico meccanismo ma per tre meccanismi indipendenti e sovrapposti, ognuno dei quali sarebbe da solo sufficiente a produrre la non-identificabilità.

VIII. Il sistema completo di teoremi di impossibilità: tassonomia strutturata

I risultati delle sezioni precedenti compongono un sistema coerente che si organizza in tre frontiere logiche distinte, con meccanismi distinti e diversi gradi di superabilità.

Prima frontiera: dal livello-1 al livello-2 (osservazione → intervento). La barriera è la non-identificabilità delle distribuzioni interventistiche da soli dati osservazionali in presenza di confondimento latente. Il meccanismo è la famiglia di modelli osservazionalmente equivalenti ma interventisticamente distinti, costruita nella Sezione II. Questa frontiera può essere *parzialmente valicata* quando si assume la conoscenza del grafo causale G : il do-calculus fornisce le condizioni necessarie e sufficienti per l'identificabilità condizionale al grafo, e l'algoritmo ID decide algebricamente se una data quantità causale è identificabile. In assenza di informazione sul grafo, la frontiera è invalicabile anche con campioni infiniti.

Seconda frontiera: struttura del grafo — dal CPDAG al DAG. Anche quando si conosce il grafo causale, i dati osservazionali identificano al più la classe di equivalenza di Markov — il CPDAG — non il DAG specifico. Gli archi non orientati del CPDAG rappresentano ambiguità irriducibili dalla sola distribuzione osservazionale. Questa frontiera può essere ridotta (ma non eliminata) con dati interventistici, come mostrano i teoremi di Eberhardt trattati nella Parte Terza. La sua struttura è di tipo combinatorio-topologico.

Terza frontiera: dal livello-2 al livello-3 (intervento → controfattuale). La barriera è la non-identificabilità della struttura di dipendenza congiunta delle variabili esogene dalle sole distribuzioni interventistiche. Il meccanismo è la costruzione della Sezione VII: funzioni strutturali con identiche distribuzioni marginali interventistiche ma diversa struttura di dipendenza congiunta delle esogene producono controfattuali distinti. Questa frontiera è *invalicabile in linea di principio* da qualsiasi quantità di esperimenti interventistici: richiederebbe la conoscenza della forma funzionale delle equazioni strutturali e della distribuzione congiunta delle esogene, quantità non accessibili dall'esterno del sistema.

La distinzione cruciale tra frontiera di principio e frontiera computazionale. La prima frontiera è una frontiera di principio in assenza di informazione sul grafo, ma diventa parzialmente valicabile quando il grafo è noto. La seconda frontiera è combinatoria: valicabile con esperimenti interventistici in numero finito bounded dai teoremi di Eberhardt. La terza frontiera è assoluta: non migliorabile con nessuna quantità di esperimenti, anche ideali.

Questa tassonomia — frontiera di principio assoluta vs. frontiera di principio condizionata alla conoscenza del grafo vs. frontiera computazionale — è lo strumento concettuale fondamentale per valutare le affermazioni sulle capacità causali dei sistemi di intelligenza artificiale, e la sua articolazione precisa è uno dei contributi principali che questa monografia offre al dibattito corrente.

IX. Estensioni: causalità in sistemi dinamici e serie temporali

I teoremi precedenti sono stati formulati in contesti statici — variabili discrete o continue senza struttura temporale esplicita. Per le applicazioni ai sistemi di intelligenza artificiale è necessario considerare le estensioni ai sistemi dinamici a tempo discreto, poiché questi sono i contesti in cui i modelli di apprendimento per rinforzo e i world models operano.

Per sistemi dinamici a tempo discreto con variabili indicizzate nel tempo $V_t = \{V_1(t), \dots, V_n(t)\}$, il modello causale strutturale diventa un sistema di equazioni strutturali ricorrenti:

$$V_i(t) := f_i(PA_{-i}(t), V_i(t-1), \dots, V_i(t-k), U_i(t))$$

dove $PA_{-i}(t)$ denota i genitori causali contemporanei di V_i al tempo t , i termini $V_i(t-j)$ modellano le dipendenze autoregressive, e $U_i(t)$ è la variabile esogena al tempo t .

Il grafo causale srotolato nel tempo (*unrolled temporal causal graph*) rappresenta ogni $V_i(t)$ come nodo distinto, con archi che codificano sia le dipendenze istantanee sia quelle ritardate. Per definizione, il grafo srotolato è aciclico (gli archi temporali vanno sempre avanti nel tempo), e la teoria dell'identificabilità si applica al grafo srotolato: un effetto causale $P(V_i(t_2) \mid \text{do}(V(t_1)=v))$ per $t_1 < t_2$ è identificabile se e solo se il sottografo srotolato tra i tempi t_1 e t_2 soddisfa le condizioni di identificabilità del do-calculus.

La causalità di Granger come causalità di livello-1. La *causalità di Granger* — introdotta da Clive Granger nel 1969 nell'analisi delle serie temporali economiche — afferma che X *causa nel senso di Granger* Y se la conoscenza dei valori passati di X migliora la previsione dei valori futuri di Y al netto dei valori passati di Y stesso:

$$X \text{ Granger-causa } Y \Leftrightarrow P(Y_t \mid Y_{-t-1}, \dots, X_{-t-1}, \dots) \neq P(Y_t \mid Y_{-t-1}, \dots)$$

Questa è una definizione di tipo predittivo-osservazionale: la causalità di Granger è una proprietà della distribuzione condizionale delle serie temporali, non degli effetti interventistici. Nella tassonomia della gerarchia causale di Pearl, si tratta di una nozione di livello-1: può coesistere con strutture causali diverse ($X \rightarrow Y$, $X \leftarrow Y$, $X \leftarrow Z \rightarrow Y$) che producono identiche implicazioni predittive, e non implica necessariamente la causalità interventistica nel senso di Pearl. Il confondimento latente dinamico — un confounder che causa sia X sia Y con ritardi diversi — produce Granger-causalità da X a Y anche in assenza di relazione causale diretta. La confusione tra causalità predittiva e causalità interventistica è una delle fonti più comuni di errori sistematici nell'analisi delle serie temporali economiche e nei sistemi di intelligenza artificiale che impiegano modelli temporali.

Il confondimento nei Markov Decision Processes. In un MDP (Markov Decision Process) con stati S_t e azioni A_t , l'agente esegue interventi $\text{do}(A_t=a)$ ad ogni passo. Se il processo è Markoviano — S_{t+1} dipende causalmente solo da (S_t, A_t) — il modello di transizione $P(S_{t+1} \mid S_t, A_t) = P(S_{t+1} \mid \text{do}(A_t=a), S_t)$ è identificabile dai dati di interazione, poiché ogni azione è per costruzione un intervento. Il confondimento può tuttavia emergere quando la politica che seleziona le azioni è correlata a caratteristiche latenti dello stato non incluse nella rappresentazione: il problema del *confounding in reinforcement learning*, attivo nell'analisi di agenti che apprendono da dimostrazioni di esperti o da dati storici invece che da esplorazione pura.

IX. Estensioni: causalità dinamica, serie temporali e sistemi adattativi

I teoremi di identificabilità, non-identificabilità e impossibilità sviluppati nelle sezioni precedenti sono stati formulati nel contesto di modelli causali strutturali statici --- sistemi in cui le variabili non sono indicizzate nel tempo e le relazioni causali sono stabili. Per le applicazioni più rilevanti ai sistemi di intelligenza artificiale e alle scienze sociali e naturali, è essenziale estendere il framework agli SCM dinamici, dove le variabili sono indicizzate nel tempo e i meccanismi causali possono cambiare. Questa sezione sviluppa la teoria causale in tre direzioni: sistemi a tempo discreto con struttura autoregressiva, sistemi a tempo continuo con equazioni differenziali stocastiche, e processi di Markov decisionali come modelli di agenti causali.

IX.1 SCM a tempo discreto: struttura formale e grafo srotolato

Definizione (SCM dinamico a tempo discreto). Un modello causale strutturale dinamico a tempo discreto con orizzonte T e una tupla $M = \langle U, V, F, P(U), T \rangle$ dove:

- $V = \{V_1, \dots, V_n\}$ è l'insieme delle variabili osservabili, con $V_t = \{V_1(t), \dots, V_n(t)\}$ il vettore delle variabili al tempo t in $\{1, \dots, T\}$

- $U = \{U_1(t), \dots, U_n(t), t \in T\}$ è l'insieme delle variabili esogene (una per variabile per tempo)

- $F = \{f_i^t : i \in \{1, \dots, n\}, t \in T\}$ è l'insieme delle equazioni strutturali, una per ogni variabile per ogni tempo

- $P(U)$ è la distribuzione congiunta delle variabili esogene

Le equazioni strutturali hanno la forma generale:

$$V_i(t) := f_i^t(PA_i(t), V_i(t-1), \dots, V_i(t-p_i), U_i(t))$$

dove $PA_i(t)$ denota i genitori causali contemporanei di V_i al tempo t (variabili $V_j(t)$ con $j \neq i$ che causano direttamente $V_i(t)$ istantaneamente), e $V_i(t-1), \dots, V_i(t-p_i)$ sono le dipendenze autoregressive di V_i con ritardo fino a p_i passi. Le variabili esogene $U_i(t)$ possono essere correlate tra loro e con $U_i(t')$ per $t' \neq t$, modellando il confondimento latente persistente nel tempo.

Il grafo srotolato nel tempo (unrolled temporal causal graph). Il grafo srotolato G_T è il DAG che rappresenta tutte le relazioni causali del sistema per tutti i passi temporali $t = 1, \dots, T$. I nodi di G_T sono le coppie (V_i, t) per ogni variabile i e ogni tempo t . Gli archi di G_T sono di due tipi:

Archi contemporanei: $(V_j, t) \rightarrow (V_i, t)$ se $V_j(t)$ è un genitore causale contemporaneo di $V_i(t)$.

Archi temporali: $(V_i, t-k) \rightarrow (V_i, t)$ per ogni ritardo $k = 1, \dots, p_i$.

Archi tra variabili diverse con ritardo: $(V_j, t-k) \rightarrow (V_i, t)$ per $V_j \neq V_i$ con lag $k \geq 1$.

Il grafo srotolato è aciclico per costruzione: gli archi temporali vanno sempre in avanti nel tempo ($t-k \rightarrow t$ con $k > 0$ garantisce l'aciclicità), e gli archi contemporanei non formano cicli se si assume che i meccanismi istantanei siano aciclici (assunzione standard negli SCM dinamici).

Esempio: VAR(1) come SCM dinamico. Il modello vettoriale autoregressivo di primo ordine (VAR(1)) per n variabili è un caso speciale di SCM dinamico lineare:

$$V(t) = A * V(t-1) + B * U(t)$$

dove A è la matrice dei coefficienti autoregressivi ($n \times n$), B è la matrice dei coefficienti degli errori ($n \times n$), e $U(t) \sim N(0, I)$ sono i termini di errore. Il grafo srotolato del VAR(1) è un DAG a livelli, con ogni $V_i(t)$ che dipende da tutti i $V_j(t-1)$ (se $A_{ij} \neq 0$) e dal proprio termine di errore. Nel VAR(1) con matrice A densa, ogni variabile causa ogni altra variabile con lag 1 --- un caso in cui il grafo srotolato è completo a ogni livello temporale.

IX.2 Identificabilità causale temporale: la dimensione dell'orizzonte

Applicazione del do-calculus al grafo srotolato. La teoria dell'identificabilità si applica direttamente al grafo srotolato. Un effetto causale $P(V_i(t_2) \mid \text{do}(V_j(t_1) = v))$ per $t_1 < t_2$ è identificabile se e solo se il sottografo srotolato $G_{\{t_1, t_2\}}$ (il sottografo indotto dai nodi tra i tempi t_1 e t_2) soddisfa le condizioni di identificabilità del do-calculus per la coppia $(V_j(t_1), V_i(t_2))$.

Teorema (identificabilità temporale e dipendenza dall'orizzonte). Sia M un SCM dinamico con confondente latente persistente U_{ij} che è causa comune di $V_i(t)$ e $V_j(t)$ per ogni t . Sia $\Delta = t_2 - t_1$ l'orizzonte temporale dell'effetto causale di interesse. Allora:

(a) Per $\Delta = 1$, l'effetto causale $P(V_i(t+1) \mid \text{do}(V_j(t) = v))$ può essere identificabile se i percorsi backdoor nel sottografo $G_{\{t, t+1\}}$ sono bloccabili tramite condizionamento su osservabili.

(b) Per $\Delta = k > 1$, il confondimento si accumula: il numero di percorsi backdoor nel sottografo $G_{\{t, t+k\}}$ cresce con k , e i percorsi backdoor di lunghezza k passano attraverso i confondenti latenti $U_{ij}(t), U_{ij}(t+1), \dots, U_{ij}(t+k-1)$. Se U è correlato nel tempo (confondimento latente persistente), i percorsi backdoor di lungo termine non sono bloccabili da condizionamento su osservabili.

(c) Per $\delta \rightarrow \infty$ (effetti stazionari a lungo termine), l'effetto causale $P(V_i(\infty) | do(V_j(0) = v))$ è in generale non identificabile in presenza di confondimento latente persistente.

Implicazione per i modelli di previsione. Il teorema sopra ha una conseguenza diretta per i modelli di previsione basati su serie temporali: i modelli che stimano $P(V_i(t+\delta) | V(1), \dots, V(t))$ identificano la distribuzione condizionale ma non l'effetto causale $P(V_i(t+\delta) | do(V_j(t) = v))$ a lungo termine in presenza di confondimento latente. La previsione condizionale e la previsione interventistica divergono a orizzonte crescente, e questa divergenza e la forma temporale del problema di identificabilità causale.

IX.3 Causalità di Granger: definizione, relazione con la causalità strutturale e limiti

Il concetto di causalità di Granger è la nozione di causalità più diffusa nelle serie temporali econometriche e nelle neuroscienze computazionali. La sua relazione con la causalità strutturale nel senso di Pearl è stata chiarita in modo definitivo da Florens e Mouchart (1985), Chamberlain (1982) e più recentemente da Eichler e Didelez (2010).

Definizione (Granger-causalità). Sia $(V_1(t), V_2(t), \dots, V_n(t))$ un processo vettoriale stazionario. La variabile V_j non Granger-causa V_i rispetto all'insieme di informazione $I_t = \{V_1(1:t), \dots, V_n(1:t)\}$ (la storia completa di tutte le variabili fino al tempo t) se:

$$V_i(t+1) \perp\!\!\!\perp V_j(1:t) \mid I_t \setminus \{V_j(1:t)\}$$

dove $I_t \setminus \{V_j(1:t)\}$ indica l'insieme di informazione con la storia di V_j rimossa. Equivalentemente, V_j non Granger-causa V_i se la storia di V_j non aggiunge potere previsivo per $V_i(t+1)$ una volta condizionato sulla storia di tutte le altre variabili.

Relazione con la causalità strutturale (Teorema di Eichler-Didelez, 2010). Le due nozioni di causalità sono in generale distinte. Specificatamente:

(a) V_j non Granger-causa V_i rispetto a I_t implica che non vi è arco diretto $V_j \rightarrow V_i$ con qualsiasi lag nel grafo causale G_T . Ma l'inverso non vale in generale: V_j può causare strutturalmente V_i e non Granger-causarla se il confondimento latente crea una dipendenza indiretta che maschera la causalità strutturale. In questo caso, V_j predice V_i non perché la causa, ma perché entrambe sono causate da un terzo fattore non osservato con effetti temporalmente sfasati.

Esempio di divergenza. Sia $U(t)$ un confondente latente con $U(t+1)$ correlato a $U(t)$ (confondimento persistente). Sia $V_j(t) = U(t) + \text{eps}_j(t)$ e sia $V_i(t+1) = U(t) + \text{eps}_i(t+1)$ (V_i dipende causalmente da U con lag 1, non da V_j). Allora V_j Granger-causa V_i (perché $\text{Cov}(V_j(t), V_i(t+1)) = \text{Cov}(U(t), U(t)) = \text{Var}(U) > 0$) ma V_j non causa strutturalmente V_i . Questo è il caso del confondimento dinamico, e rappresenta la forma temporale del problema di identificabilità causale.

Implicazione per la scoperta causale da serie temporali. La causalità di Granger non identifica la struttura causale strutturale in presenza di confondimento latente. Algoritmi come il Granger causality test, la Transfer Entropy, o il PCMC (Peters et al., 2017) - che estende l'algoritmo PC alle serie temporali - producono grafi di Granger-causalità che possono divergere dal vero grafo causale strutturale in presenza di confondenti latenti. Il PCMC migliora rispetto al Granger test standard controllando per le dipendenze indirette, ma non risolve il problema del confondimento latente.

IX.4 SCM a tempo continuo: equazioni differenziali stocastiche e identificabilità

Per sistemi fisici, biologici ed economici modellati in tempo continuo, il framework degli SCM si estende alle equazioni differenziali stocastiche (SDE). Il vantaggio del tempo continuo è che elimina l'arbitrarietà del passo di discretizzazione e permette di modellare i processi causali alla loro scala temporale naturale.

Definizione (SCM a tempo continuo). Un SCM a tempo continuo è un sistema di SDE della forma:

$$dV_i(t) = f_i(V(t), PA_i(t), t) dt + \sigma_i(V(t), t) dW_i(t)$$

dove f_i è il drift causale della variabile V_i (determinato dai suoi genitori causali $PA_i(t)$ e da tutti i $V(t)$), σ_i è il coefficiente di diffusione, e $W_i(t)$ sono moti Browniani indipendenti. Il grafo causale G del sistema è determinato dalle dipendenze del drift: $V_j \rightarrow V_i$ in G se e solo se $df_i/dV_j \neq 0$ per qualche configurazione delle variabili.

Identificabilità causale in sistemi a tempo continuo. Mooij, Janzing e Scholkopf (2013) hanno dimostrato che in sistemi a tempo continuo con rumore additivo (σ_i costante), il problema dell'identificabilità causale ha proprietà diverse rispetto al caso discreto. In particolare, sotto condizioni di regolarità del drift f_i e indipendenza dei rumori, la distribuzione stazionaria $P(V)$ determina il grafo causale sotto assunzioni funzionali più deboli che nel caso discreto. Intuitivamente, il tempo continuo permette di sfruttare le

proprietà della distribuzione invariante (stazionaria) per distinguere la direzione causale dal comportamento del sistema vicino alla stazionarietà.

Limite fondamentale: i processi a equilibrio. Per sistemi che raggiungono un equilibrio --- come molti sistemi biologici e fisici --- la distribuzione osservata è la distribuzione stazionaria, che è determinata dall'equazione di Fokker-Planck. In questo regime, la distribuzione stazionaria non porta informazione sulla direzione temporale dei meccanismi causali, e il problema della identificabilità causale è identico al caso statico. In pratica, per sistemi al loro equilibrio, la causalità temporale non aggiunge potere identificativo rispetto ai metodi statici.

IX.5 Processi di Markov decisionali come SCM dinamici: il problema del confondimento in RL

I sistemi di apprendimento per rinforzo operano in ambienti modellati come MDP (Markov Decision Process), dove un agente osserva stati S_t , sceglie azioni A_t secondo una politica π , e riceve ricompense R_t . La struttura di un MDP è formalmente equivalente a un SCM dinamico in cui le azioni dell'agente sono interventi causali, e questa equivalenza ha implicazioni profonde per la comprensione delle capacità e dei limiti degli agenti causali.

Struttura del MDP come SCM. In un MDP, le equazioni strutturali sono:

$$A_t := \pi(S_t, H_t) \text{ (politica dell'agente)}$$

$$S_{t+1} := f(S_t, A_t, U_t) \text{ (transizione di stato)}$$

$$R_t := r(S_t, A_t, V_t) \text{ (funzione di ricompensa)}$$

dove H_t è la storia passata (per politiche non Markoviane), e U_t, V_t sono variabili esogene. Per la proprietà di Markov degli MDP, S_{t+1} dipende causalmente solo da S_t e A_t - non dalla storia. In questo caso, l'azione A_t è un intervento causale in senso stretto: l'agente rompe il meccanismo naturale di selezione di A_t (che nella politica behaviour sarebbe determinato da S_t e dalla storia) e impone il valore $A_t = a$.

Identificabilità del modello di transizione. Per MDP Markoviani in cui la politica è esplorativa (ogni azione ha probabilità positiva in ogni stato), il modello di transizione $P(S_{t+1} \mid S_t, A_t)$ è identico all'effetto causale $P(S_{t+1} \mid \text{do}(A_t = a), S_t)$. Questo è il caso privilegiato in cui l'apprendimento per rinforzo è genuinamente causale: l'agente apprende il modello causale di transizione attraverso l'esplorazione interventistica dell'ambiente.

Confondimento in RL: l'off-policy evaluation problem. Il problema del confondimento emerge quando si vuole valutare una nuova politica π_{new} usando dati raccolti sotto una politica di comportamento diversa $\pi_{\text{behaviour}}$. Formalmente, si vuole calcolare:

$$V(\pi_{\text{new}}) = E_{\{\pi_{\text{new}}\}}[\sum_t \gamma^t * R_t]$$

usando solo i dati $\{(S_t, A_t, R_t, S_{t+1})\}$ raccolti con $\pi_{\text{behaviour}}$. Questo è il problema dell'off-policy evaluation (OPE). Il confondimento emerge perché la politica $\pi_{\text{behaviour}}$ seleziona A_t in funzione di S_t : la distribuzione dei dati osservazionali è $P(S_t, A_t) = P(A_t | S_t) * P(S_t)$, non la distribuzione di interesse $P_{\{\pi_{\text{new}}\}}(S_t, A_t)$. Sotto l'assunzione di copertura (overlap): $\pi_{\text{new}}(a | s) > 0$ implica $\pi_{\text{behaviour}}(a | s) > 0$, il bias di confondimento può essere corretto con importance weighting:

$$\hat{V}(\pi_{\text{new}}) = E_{\{\pi_{\text{behaviour}}\}}[\sum_t \gamma^t * (\pi_{\text{new}}(A_t | S_t) / \pi_{\text{behaviour}}(A_t | S_t)) * R_t]$$

Questo è l'estimatore importance sampling per OPE, ed è formalmente equivalente all'aggiustamento causale do-calculus nel caso dei dati interventistici con politica nota. Quando la politica di comportamento è ignota (dati osservazionali da un agente umano con politica non modellata), l'assunzione di copertura non può essere verificata e il bias di confondimento non può essere corretto senza assunzioni aggiuntive.

Il confondimento latente in RL: stati parzialmente osservati. In MDP parzialmente osservati (POMDP), l'agente osserva solo $O_t = g(S_t, N_t)$ (una funzione rumorosa dello stato), non lo stato completo S_t . In questo caso, la politica $\pi(O_t)$ è correlata con la parte dello stato non osservata S_t^{hidden} , creando un confondimento latente tra l'azione e la transizione di stato. Formalmente, il POMDP è equivalente a un SCM con variabile latente S_t^{hidden} che è causa comune di A_t (attraverso O_t e la politica) e di S_{t+1} (attraverso la transizione di stato). La struttura causale del POMDP è quindi identica alla struttura del problema di non-identificabilità con variabili latenti della Parte Seconda: il confondimento latente rende l'effetto causale delle azioni sul futuro non identificabile in generale senza l'assunzione di osservabilità completa.

IX.6 Stabilità causale e cambi di regime: il problema della non-stazionarietà

I teoremi di impossibilità sviluppati nelle sezioni precedenti assumono la stazionarietà del meccanismo causale: le equazioni strutturali f_i sono costanti nel tempo. In molte applicazioni rilevanti --- sistemi economici in recessione, sistemi biologici in risposta a perturbazioni, sistemi sociali durante le crisi - i meccanismi causali cambiano nel tempo. Questo problema è noto come distributional shift o covariate shift nella letteratura di machine learning, e come instabilità dei parametri strutturali nella letteratura econometrica.

Definizione (invarianza causale, Peters, Mooij, Janzing, Scholkopf, 2016). Un meccanismo causale $V_i := f_i(PA_i, U_i)$ si dice causalmente invariante tra gli ambienti e in $E = \{e_1, e_2, \dots\}$ se la distribuzione condizionale $P_e(V_i \mid PA_i)$ è la stessa per tutti e in E . La proprietà di invarianza causale distingue i meccanismi causali (le equazioni strutturali, che sono invarianti per ipotesi) dalle distribuzioni marginali (che possono cambiare tra ambienti).

Teorema (invarianza causale e identificabilità tra ambienti). Sotto l'assunzione di invarianza causale, è possibile identificare i meccanismi causali (e quindi la struttura del grafo) sfruttando le differenze tra ambienti: se una distribuzione condizionale $P_e(V_i \mid Z)$ varia con l'ambiente e allora Z non è il set di genitori causali di V_i , ma include variabili non causalmente rilevanti o mancano dei genitori causali reali. L'algoritmo ICP (Invariant Causal Prediction) di Peters et al. (2016) sfrutta questo principio per identificare i genitori causali di una variabile di interesse tra ambienti multipli, senza richiedere la randomizzazione degli interventi.

Questo risultato apre una direzione di ricerca importante per l'inferenza causale in sistemi non stazionari: invece di cercare di identificare la struttura causale da un singolo dataset osservazionale (problema in generale impossibile), si sfruttano le variazioni tra ambienti come una forma di quasi-intervento. Il collegamento con i metodi quasi-sperimentali della sezione precedente è diretto: ogni cambio di regime è un quasi-esperimento naturale sulle variabili che cambiano, e l'invarianza causale delle equazioni strutturali è l'equivalente dell'assunzione di esogeneità strumentale nel metodo IV.

X. Implicazioni per il metodo scientifico: ridefinire il ruolo dei dati

Le implicazioni del sistema di teoremi di impossibilità presentato in questa Parte si estendono ben oltre la teoria formale. Il quadro ridefinisce il ruolo dei dati nell'inferenza causale in un modo che contrasta con il mito dominante — diffuso tanto nel dibattito sull'intelligenza artificiale quanto nella scienza dei dati applicata — secondo cui l'accumulo di dati osservazionali sarebbe sufficiente per rispondere a domande causali.

I dati osservazionali non sono semplicemente insufficienti per rispondere a domande causali: in presenza di confondimento latente, sono logicamente inadatti allo scopo — anche in campioni infiniti, anche con algoritmi privi di errore computazionale. L'inferenza causale da dati osservazionali è possibile esclusivamente quando si assume una struttura causale qualitativa del fenomeno — un grafo G derivato dalla teoria del dominio, da esperimenti precedenti, o da assunzioni di identificabilità strutturalmente motivate — e quando quel grafo soddisfa le condizioni di identificabilità per le quantità di interesse.

Questa ridefinizione del ruolo dei dati non è scettica: non nega la possibilità dell'inferenza causale. Afferma che l'inferenza causale è un'impresa mista — che combina dati quantitativi con conoscenza qualitativa della struttura — e che la qualità della struttura assunta determina la qualità dell'inferenza causale in modo non recuperabile da nessuna quantità di dati aggiuntivi. L'idea che sistemi di apprendimento automatico addestrati su enormi corpora di dati osservazionali possano apprendere strutture causali del mondo è soggetta a una critica di principio: apprendere la struttura causale richiede dati di tipo qualitativamente diverso — interventistici — e anche con questi ultimi i controfattuali rimangono parzialmente fuori portata, per ragioni che i teoremi di questa Parte hanno reso formalmente precise.

Parte Terza: Classi di Equivalenza di Markov e la Topologia dello Spazio dei Modelli Causali

I. Il problema dell'identità causale nella sua dimensione topologica

La Parte Seconda ha dimostrato che la distribuzione osservazionale $P(V)$, da sola, non determina la distribuzione post-intervento $P(V \mid \text{do}(X=x))$ in presenza di confondimento latente. Ma la natura e la misura dell'ambiguità residua merita di essere posta in termini ancora più fondamentali: quanti grafi causali distinti sono compatibili con una data distribuzione $P(V)$, anche nella situazione ideale di sufficienza causale — quando si è certi che tutte le variabili causalmente rilevanti siano osservate e non vi siano variabili latenti? La risposta è contenuta nella teoria delle classi di equivalenza di Markov, sviluppata da Thomas Verma e Judea Pearl nei lavori del 1990–1992 e poi sistematizzata da Christopher Meek, David Chickering e dal gruppo di Andersson, Madigan e Perlman nella seconda metà degli anni Novanta.

Il punto di partenza è una domanda di geometria degli spazi di modelli: nella varietà di tutti i DAG su un insieme fisso di variabili $V = \{V_1, \dots, V_n\}$, due DAG G_1 e G_2 sono osservazionalmente indistinguibili quando nessun esperimento passivo — nessuna raccolta di dati senza intervento, per quanti dati si raccolgano — può distinguerli. La relazione di indistinguibilità osservazionale è una relazione di equivalenza, e le classi di equivalenza che essa genera — le classi di Markov — partizionano lo spazio dei DAG in famiglie di grafi con identiche implicazioni osservazionali. Questa topologia ha conseguenze dirette per la pratica: un algoritmo di scoperta causale che operi su dati osservazionali, per quanto corretto e consistente, non può identificare il DAG causale ma soltanto la sua classe di equivalenza. La struttura della classe — quanti grafi contiene, quali archi sono orientati invariabilmente e quali no, quanto differiscono le predizioni interventistiche dei diversi membri — determina l'ambiguità residua irriducibile di qualsiasi procedura di scoperta causale osservazionale.

Prima di procedere alla teoria formale, vale la pena chiarire perché la sufficienza causale sia raramente soddisfatta nella pratica, e cosa cambia quando viene rilasciata. Sotto sufficienza causale, il limite dell'inferenza osservazionale è il CPDAG. In presenza di variabili latenti, il limite si arretra al PAG (Partial Ancestral Graph), una struttura ancora più ambigua che incapsula sia l'incertezza sulle direzioni degli archi sia l'incertezza sulla presenza di

confondimento latente. La Sezione VII esplora la teoria dei MAG e dei PAG, rilevante per ogni applicazione reale in cui l'elenco completo delle variabili causalmente rilevanti non è noto.

II. Fondamenti della d-separazione: il linguaggio delle indipendenze condizionali

La d-separazione è il criterio con cui si leggono le indipendenze condizionali direttamente dalla struttura grafica del DAG, senza necessità di calcolare le distribuzioni di probabilità. Essa costituisce il fondamento dell'intera teoria dell'equivalenza di Markov, e la sua comprensione approfondita è il prerequisito per ogni applicazione successiva.

Definizione (percorso in un DAG). Un percorso π tra A e B in un DAG G è una sequenza di nodi $V_1 = A, V_2, \dots, V_k = B$ tali che ogni coppia consecutiva (V_i, V_{i+1}) sia adiacente in G — esiste un arco tra V_i e V_{i+1} in qualche direzione. Il percorso è diretto se ogni arco è orientato nella direzione della sequenza ($V_i \rightarrow V_{i+1}$); è misto se contiene almeno un arco in direzione opposta.

Definizione (ruoli dei nodi interni). In un percorso $\pi = (V_1, \dots, V_k)$, un nodo interno V_i con $1 < i < k$ è un *collider* sul percorso se entrambi gli archi adiacenti puntano verso di esso: $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$. È un *noncollider* in tutti gli altri casi: catena ($V_{i-1} \rightarrow V_i \rightarrow V_{i+1}$ o la sua inversa), biforcazione ($V_{i-1} \leftarrow V_i \rightarrow V_{i+1}$).

Definizione (d-separazione). Un percorso π tra A e B è bloccato da un insieme di nodi C se almeno una delle seguenti condizioni vale: (i) il percorso contiene un noncollider M con $M \in C$; (ii) il percorso contiene un collider M con $M \notin C$ e nessun discendente di M appartiene a C . Due nodi A e B sono d-separati da C in G , scritto $A \perp_d B \mid C$ in G , se ogni percorso tra A e B nel grafo è bloccato da C .

La logica del blocco dei noncollider è intuitiva: $M \in C$ interrompe il flusso di informazione perché M è un mediatore (catena) o un confondente (biforcazione) la cui variabilità trasmette la correlazione tra A e B ; fissando M si taglia questa trasmissione. La logica del collider è controintuitiva ma strutturalmente fondamentale: il collider $M \notin C$ lascia il percorso bloccato — non condizionare sul collider non trasmette correlazione tra le sue cause — mentre il collider $M \in C$ apre il percorso — condizionare sul collider crea una dipendenza tra le sue cause attraverso il collider bias. Questa asimmetria — il condizionamento blocca i noncollider ma apre i collider — è il cuore della d-separazione e la ragione per cui la terza risoluzione del paradosso di Simpson è la meno intuitiva delle tre.

L'algoritmo di Bayes-Ball. Un algoritmo efficiente per decidere la d-separazione è l'algoritmo di Bayes-Ball di Shachter (1998), che visita il grafo simulando il rimbalzo di palline da A verso B secondo le regole di blocco e attivazione, determinando in tempo $O(|V| + |E|)$ se qualche pallina raggiunge B partendo da A attraverso un percorso non bloccato da C .

Il teorema di Markov e la fedeltà. La connessione tra d-separazione grafica e indipendenza statistica è garantita da due proprietà: la condizione di Markov (ogni d-separazione nel grafo implica l'indipendenza statistica nella distribuzione generata dal modello) e la condizione di

fedeltà (ogni indipendenza statistica nella distribuzione è implicata da una d-separazione nel grafo — non vi sono indipendenze accidentali). La condizione di fedeltà può fallire per cancellazioni parametriche, e la Sezione VIII analizza in dettaglio questo fenomeno.

III. Il teorema di Verma: caratterizzazione completa dell'equivalenza di Markov

Definizione (equivalenza di Markov). Due DAG G_1 e G_2 su V sono Markov-equivalenti — scritto $G_1 \equiv_M G_2$ — se per ogni distribuzione di probabilità $P(V)$ e per ogni tripla (A, B, C) con $A, B \in V$ e $C \subseteq V \setminus \{A, B\}$:

$$(A \perp B \mid C)\{G_1\} \Leftrightarrow (A \perp B \mid C)\{G_2\}$$

In altri termini, G_1 e G_2 sono Markov-equivalenti quando hanno esattamente le stesse implicazioni di d-separazione: nessun test di indipendenza condizionale può distinguerli, qualunque sia la distribuzione osservazionale generata dall'uno o dall'altro.

Due DAG Markov-equivalenti possono generare distribuzioni quantitativamente diverse (con parametri distinti) pur condividendo esattamente la struttura delle dipendenze condizionali. L'equivalenza di Markov è definita sulle implicazioni di d-separazione, non sulle distribuzioni stesse.

Definizione (scheletro). Lo *scheletro* di un DAG G — denotato $sk(G)$ — è il grafo non orientato H ottenuto ignorando le direzioni degli archi: H contiene un arco non orientato tra V_i e V_j se e solo se G contiene un arco orientato $V_i \rightarrow V_j$ o $V_j \rightarrow V_i$. Lo scheletro cattura la struttura di adiacenza del DAG, indipendente dalla direzione degli archi.

Definizione (v-struttura). Una *v-struttura* — anche chiamata collider aperto o struttura a forcella capovolta — in un DAG G è una tripla ordinata di nodi (A, C, B) tale che: (i) $A \rightarrow C$ e $B \rightarrow C$ sono archi in G (entrambi gli archi entrano in C); (ii) A e B non sono adiacenti in G . La condizione di non adiacenza è essenziale: senza di essa, la tripla (A, C, B) con $A \rightarrow C \leftarrow B$ in cui A e B sono adiacenti sarebbe un *protected collider* e non avrebbe la stessa implicazione di d-separazione.

Teorema (Verma e Pearl, 1990 — caratterizzazione dell'equivalenza di Markov). Due DAG G_1 e G_2 su V sono Markov-equivalenti se e soltanto se: (i) $sk(G_1) = sk(G_2)$ — hanno lo stesso scheletro; (ii) G_1 e G_2 hanno le stesse v-strutture — per ogni tripla (A, C, B) , (A, C, B) è una v-struttura in G_1 se e solo se lo è in G_2 .

Dimostrazione. La dimostrazione procede in due direzioni.

Direzione ($\Rightarrow$): l'equivalenza implica stesso scheletro e stesse v-strutture.

Stesso scheletro. Supponiamo $G_1 \equiv_M G_2$ e supponiamo per assurdo che abbiano scheletri diversi. Allora esiste una coppia (A, B) adiacente in G_1 ma non in G_2 . In G_2 , A e B non sono adiacenti. Per il teorema dell'insieme di d-separazione — in ogni DAG, due nodi non adiacenti sono d-separati da qualche insieme — esiste C tale che $(A \perp B \mid C)\{G_2\}$.

L'insieme che d-separa A da B in G_2 è un sottoinsieme dei genitori di A o dei genitori di B, come segue dall'algoritmo di Bayes-Ball.

Ma in G_1 , A e B sono adiacenti: un arco diretto tra due nodi non può essere bloccato da alcun insieme di condizionamento, poiché crea una dipendenza che nessun condizionamento elimina, per la proprietà di Markov in senso stretto. Quindi per ogni insieme C, $(A \not\perp_d B \mid C)\{G_1\}$, il che contraddice $G_1 \equiv_M G_2$. Dunque $sk(G_1) = sk(G_2)$.

Stesse v-strutture. Supponiamo ora che G_1 e G_2 abbiano lo stesso scheletro ma v-strutture diverse. Esiste allora una tripla (A, C, B) che è v-struttura in G_1 — con $A \rightarrow C \leftarrow B$ e A non adiacente a B — ma non v-struttura in G_2 . Poiché i due grafi hanno lo stesso scheletro, gli archi A-C e B-C esistono in G_2 ; poiché (A, C, B) non è v-struttura in G_2 , almeno un arco è invertito: diciamo $A \leftarrow C$ in G_2 (le situazioni $A \leftarrow C \leftarrow B$, $A \leftarrow C \rightarrow B$, $A \rightarrow C \rightarrow B$ sono analoghe per simmetria).

In G_1 , la tripla (A, C, B) è una v-struttura: per le regole della d-separazione, A e B sono d-separati dall'insieme vuoto $\emptyset$ — il percorso $A \rightarrow C \leftarrow B$ ha C come collider, e un collider blocca il percorso quando non è condizionato — ma *non* sono d-separati da $\{C\}$ — condizionare sul collider C apre il percorso. Quindi:

$$(A \perp_d B \mid \emptyset)\{G_1\} \text{ ma } (A \not\perp_d B \mid \{C\})\{G_1\}$$

In G_2 , con la struttura $A \leftarrow C \rightarrow B$ (biforcazione, C come causa comune), il percorso $A \leftarrow C \rightarrow B$ ha C come noncollider. Le regole della d-separazione danno:

$$(A \not\perp_d B \mid \emptyset)\{G_2\} \text{ e } (A \perp_d B \mid \{C\})\{G_2\}$$

Le strutture di d-separazione di G_1 e G_2 sono opposte: nella tripla (A, C, B), G_1 implica $A \perp_d B \mid \emptyset$ ma non $A \perp_d B \mid \{C\}$, mentre G_2 implica il contrario. Questo contraddice $G_1 \equiv_M G_2$. $\square$

Direzione ($\Leftarrow$): stesso scheletro e stesse v-strutture implicano equivalenza.

Se G_1 e G_2 hanno lo stesso scheletro e le stesse v-strutture, dimostriamo che hanno le stesse implicazioni di d-separazione.

Sia π un percorso qualunque tra A e B in G_1 (che è lo stesso percorso in G_2 per la coincidenza degli scheletri). Vogliamo mostrare che π è bloccato da C in G_1 se e solo se è bloccato da C in G_2 . Per ogni nodo interno V_i sul percorso π (con $1 < i < k$):

Caso V_i è un collider in G_1 . Allora gli archi $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$ sono entrambi entranti in V_i . Se V_{i-1} e V_{i+1} non sono adiacenti, allora (V_{i-1}, V_i, V_{i+1}) è una v-struttura in G_1 . Per l'identità delle v-strutture, è anche una v-struttura in G_2 : gli archi sono orientati allo stesso modo ($V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$ anche in G_2). Dunque V_i è collider anche in G_2 sullo stesso percorso. Le condizioni di blocco per il collider dipendono da: (a) se $V_i \in C$, il collider è attivato; (b) se qualche discendente di $V_i \in C$, il collider è attivato. La condizione (a) è identica in G_1 e G_2 . La condizione (b) dipende dall'insieme dei discendenti di V_i , che può differire tra G_1 e G_2 per gli archi uscenti da V_i . Tuttavia, Chickering (1995) dimostra che, per la struttura complessiva dei percorsi non bloccati compatibili con le stesse v-strutture, questa potenziale differenza non

produce differenze nella relazione di d-separazione finale tra A e B : le strutture di d-separazione coincidono a livello globale.

Caso V_i è un noncollider in G_1 . Le possibilità sono la catena ($V_{i-1} \rightarrow V_i \rightarrow V_{i+1}$ o la sua inversa) o la biforcazione ($V_{i-1} \leftarrow V_i \rightarrow V_{i+1}$). In nessuno di questi casi si ha una v-struttura in V_i . L'identità delle v-strutture garantisce che se V_i è noncollider su questo percorso in G_1 , non può essere collider in G_2 sullo stesso percorso: altrimenti la tripla (V_{i-1}, V_i, V_{i+1}) sarebbe v-struttura in G_2 ma non in G_1 , poiché V_{i-1} e V_{i+1} non sono adiacenti (lo scheletro è identico e nel percorso non vi è arco tra V_{i-1} e V_{i+1}). La condizione di blocco per i noncollider — $V_i \in C$ — dipende solo dalla presenza di V_i nel set di condizionamento, non dalla direzione degli archi. Dunque il blocco del percorso in V_i è identico in G_1 e G_2 .

Poiché le condizioni di blocco sono identiche per ogni nodo interno di ogni percorso, le strutture di d-separazione di G_1 e G_2 coincidono, e $G_1 \equiv_M G_2$. $\square$

IV. Il CPDAG: la rappresentazione canonica della classe di equivalenza

Il teorema di Verma riduce ogni classe di equivalenza di Markov a due invarianti combinatoriali: lo scheletro (il grafo non orientato degli archi) e le v-strutture (le triple collider con nodi estremi non adiacenti). Il CPDAG (Complete Partially Directed Acyclic Graph) codifica entrambe in un unico oggetto canonico che serve da rappresentante privilegiato della classe.

Definizione (CPDAG). Il CPDAG di una classe di equivalenza $[G]$ è il grafo parzialmente orientato C su V tale che: C contiene l'arco orientato $V_i \rightarrow V_j$ se e soltanto se tutti i DAG della classe hanno l'arco $V_i \rightarrow V_j$ — l'arco è *orientato invariante*; C contiene l'arco non orientato $V_i - V_j$ se e soltanto se esistono DAG della classe con direzioni opposte per quell'arco — l'arco è *genuinamente ambiguo*.

Costruzione del CPDAG — le regole di Meek. Dato un DAG G , il suo CPDAG si costruisce in tre fasi: (1) si estrae lo scheletro $sk(G)$; (2) si orientano tutte le v-strutture; (3) si applicano iterativamente le quattro regole di Meek fino a convergenza, orientando ogni arco aggiuntivo il cui verso è forzato dalla struttura già presente. Le regole di Meek esprimono le uniche situazioni in cui un arco non ancora orientato deve assumere una direzione specifica per non violare le condizioni di aciclicità o non introdurre v-strutture assenti nel DAG originale.

Regola R1: se $A \rightarrow B - C$ con A non adiacente a C , si orienta $B \rightarrow C$. Ragione: orientare $C \rightarrow B$ creerebbe la v-struttura (A, B, C) assente nel grafo originale.

Regola R2: se $A \rightarrow B \rightarrow C$ e $A - C$, si orienta $A \rightarrow C$. Ragione: orientare $C \rightarrow A$ creerebbe il ciclo diretto $A \rightarrow B \rightarrow C \rightarrow A$.

Regola R3: se $A \rightarrow C \leftarrow B$ con $A - D, B - D, A$ non adiacente a B , e $D - C$, si orienta $D \rightarrow C$. Ragione: ogni altra orientazione di $D - C$ creerebbe una v-struttura non presente o una violazione delle regole precedenti in combinazione con la struttura esistente.

Regola R4: configurazioni a quattro nodi che forzano orientazioni particolari per preservare la consistenza delle v-strutture e l'aciclicità in presenza di triple sovrapposte.

Teorema (completezza delle regole di Meek, 1995). Le regole R1–R4 applicate iterativamente terminano in tempo polinomiale e orientano esattamente gli archi il cui verso è invariante in tutti i DAG della classe. Ogni arco rimasto non orientato al termine dell'iterazione è genuinamente ambiguo — esistono effettivamente due DAG nella classe con orientazioni opposte per quell'arco.

L'unicità del CPDAG come rappresentante canonico garantisce che ogni classe di equivalenza abbia esattamente un CPDAG, e che il CPDAG sia costruibile da qualsiasi DAG della classe con la procedura sopra descritta.

V. Dimensioni delle classi di equivalenza e quantificazione dell'ambiguità

Conteggio dei DAG etichettati. Il numero $a(n)$ di DAG etichettati su n nodi soddisfa la ricorrenza di Robinson:

$$a(n) = \sum_{k=1}^n (-1)^{k+1} \cdot C(n, k) \cdot 2^{k(n-k)} \cdot a(n-k), \text{ con } a(0) = 1$$

I valori iniziali sono: $a(1)=1$, $a(2)=3$, $a(3)=25$, $a(4)=543$, $a(5)=29281$, $a(6)=3781503$. La crescita è super-esponenziale.

Classi di equivalenza di Markov. Il numero $e(n)$ di CPDAG distinti su n nodi è: $e(1)=1$, $e(2)=2$, $e(3)=11$, $e(4)=185$, $e(5)=8782$. Il rapporto $a(n)/e(n)$ — che misura la dimensione media delle classi — cresce con n , indicando che classi più grandi diventano tipiche all'aumentare della dimensione del sistema.

Risultati per grafi sparsi (Gillispie e Perlman, 2002). Per DAG con grado medio d , la dimensione media delle classi scala come $O(n^{d-1})$: per $d=2$, $O(n)$; per $d=3$, $O(n^2)$; per $d=4$, $O(n^3)$. Per $n=100$ variabili e grado medio 3, le classi contengono tipicamente $O(10^4)$ DAG. Ancora più significativo è il risultato sulla proporzione di archi non orientati nel CPDAG: per grafi uniformemente casuali di densità fissa, questa proporzione converge a una costante strettamente positiva con il crescere di n — l'ambiguità strutturale non svanisce con la dimensione del sistema, bensì diventa una caratteristica tipica e stabile.

Il significato interventistico dell'ambiguità. Due DAG nella stessa classe di equivalenza possono produrre distribuzioni post-intervento radicalmente diverse. Per il caso bivariato non orientato $A \text{ — } B$: se $A \rightarrow B$ è il DAG vero, l'intervento $\text{do}(A=a)$ produce un effetto su B ; se $B \rightarrow A$ è il DAG vero, lo stesso intervento non ha effetto su B . Le distribuzioni osservazionali sono identiche, le distribuzioni post-intervento sono opposte. L'ambiguità del CPDAG si traduce direttamente in incapacità decisionale concreta: un sistema che deve scegliere tra le due politiche interventistiche sulla base di soli dati osservazionali è strutturalmente incapace di farlo, non per mancanza di dati ma per struttura logica del problema.

VI. La riduzione dell'ambiguità con dati interventistici: teoremi di Eberhardt

Definizione (I-equivalenza). Due DAG G_1 e G_2 sono I-equivalenti rispetto a una famiglia di interventi $\mathbb{I}$ se per ogni intervento $I \in \mathbb{I}$ e ogni tripla (A, B, C) :

$$(A \perp_d B \mid C) \text{ nel grafo mutilato } G_1^{\wedge I} \Leftrightarrow (A \perp_d B \mid C) \text{ nel grafo mutilato } G_2^{\wedge I}$$

La I-equivalenza è più fine dell'equivalenza di Markov: aggiungere dati interventistici può distinguere DAG che erano osservazionalmente indistinguibili, riducendo le classi di equivalenza a sottoclassi più piccole.

Teorema (Eberhardt, 2007). (a) *Upper bound*: $n-1$ interventi adattativi a singolo nodo sono sempre sufficienti per identificare il DAG causale da qualsiasi classe di equivalenza su n nodi. (b) *Lower bound*: esistono classi di equivalenza per cui $\lfloor n/2 \rfloor$ interventi sono necessari anche con strategia adattativa ottimale.

L'upper bound $n-1$ si dimostra per induzione: ogni intervento su un singolo nodo X divide il grafo in due parti (i nodi con X come causa e i nodi con X come effetto), e $n-1$ interventi sul sistema completo orientano tutti gli archi rimanenti. Il lower bound $\lfloor n/2 \rfloor$ si dimostra costruendo classi di equivalenza particolarmente simmetriche in cui ogni intervento su un singolo nodo risolve l'ambiguità di al più due archi. Il gap $n-1$ vs. $\lfloor n/2 \rfloor$ corrisponde all'efficienza della strategia adattativa: il fattore 2 di risparmio si ottiene scegliendo ottimalmente l'intervento successivo sulla base del CPDAG residuo aggiornato dopo ogni esperimento.

Teorema (Hauser e Bühlmann, 2012 — interventi ottimali). L'insieme minimo di interventi per identificare il DAG da un CPDAG noto è determinato dagli archi protetti del CPDAG. Algoritmi greedy che massimizzano il numero di archi orientati per intervento hanno complessità vicina al lower bound di Eberhardt e sono computazionalmente efficienti, con aggiornamento del CPDAG in tempo $O(n^3)$.

La classe di equivalenza interventistica (IMEC). Hauser e Bühlmann introducono la nozione di Interventional Markov Equivalence Class: la famiglia di DAG non distinguibili da una combinazione di dati osservazionali e da una famiglia fissa di interventi $\mathbb{I}$. Il rappresentante canonico è l'I-CPDAG, che orienta gli archi aggiuntivamente identificabili dagli interventi rispetto al CPDAG osservazionale. Questa struttura fornisce il fondamento formale per il disegno ottimale di esperimenti causali sequenziali — un problema con rilevanza diretta nella biologia molecolare (dove ogni esperimento di knockout genico è costoso), nell'econometria sperimentale e nella progettazione di agenti autonomi che esplorano attivamente il proprio ambiente.

V. Dimensioni delle classi di equivalenza: calcolo asintotico di Gillispie-Perlman

La teoria delle classi di equivalenza di Markov stabilisce che ogni distribuzione osservazionale determina al più il CPDAG, non il DAG. Per valutare il peso in pratica di questa ambiguità è necessario quantificare le dimensioni tipiche delle classi: quanti DAG distinti appartengono alla stessa classe? Gillispie e Perlman (2002) hanno fornito la risposta formale più completa a questa domanda, derivando asintotici precisi per la dimensione media delle classi in funzione del numero di nodi e della sparsità del grafo.

V.1 Il conteggio esatto dei DAG etichettati: la formula ricorrente di Robinson

Il numero $a(n)$ di DAG etichettati distinti su n nodi - dove "etichettati" significa che i nodi hanno identità distinte e i grafi che differiscono solo per la denominazione dei nodi sono considerati distinti - soddisfa la ricorrenza di Robinson (1973):

$$a(n) = \sum_{k=1}^n (-1)^{k+1} * C(n,k) * 2^{k(n-k)} * a(n-k), \quad a(0) = 1$$

La derivazione di questa formula si basa sul principio di inclusione-esclusione applicato al numero di nodi sorgente (nodi senza genitori nel DAG). Ogni DAG su n nodi ha almeno un nodo sorgente; il termine k -esimo della somma conta i DAG con esattamente k nodi sorgente specifici, con il fattore $2^{k(n-k)}$ che conta le scelte degli archi dai k nodi sorgente verso i rimanenti $n-k$ nodi.

I valori iniziali sono:

$$a(1) = 1, a(2) = 3, a(3) = 25, a(4) = 543, a(5) = 29281, a(6) = 3781503.$$

La crescita è super-esponenziale: $a(n) = \Theta(n! * 2^{n(n-1)/2} / C)$ per una costante C , come dimostrato da Robinson (1977) per via asintotica.

Derivazione della ricorrenza. Sia S un sottoinsieme di $\{1, \dots, n\}$ di cardinalità k . Un DAG su $\{1, \dots, n\}$ ha S come insieme di nodi sorgente (nodi senza genitori) se e solo se: (a) ogni nodo in S non ha genitori; (b) ogni nodo in $V \setminus S$ ha almeno un genitore in S o in $V \setminus S$. Poiché i nodi in S non hanno genitori, gli archi da S verso $V \setminus S$ sono completamente liberi: $2^{k(n-k)}$ scelte. I nodi in $V \setminus S$ formano un DAG su $n-k$ nodi: $a(n-k)$ scelte.

Applicando il principio di inclusione-esclusione su tutti i sottoinsiemi S di taglia k , con il segno alternato per escludere i DAG con insieme di sorgenti più grande di S :

$$a(n) = \sum_{k=1}^n (-1)^{k+1} * C(n,k) * 2^{k(n-k)} * a(n-k)$$

I valori calcolati confermano la super-crescita: per $n=10$, $a(10)$ è dell'ordine di 10^{18} .

V.2 Il conteggio delle classi di equivalenza di Markov

Il numero $e(n)$ di CPDAG distinti su n nodi - equivalentemente, il numero di classi di equivalenza di Markov distinte - è stato tabulato da Gillispie e Perlman (2002) per piccoli n :

$e(1) = 1, e(2) = 2, e(3) = 11, e(4) = 185, e(5) = 8782.$

La crescita di $e(n)$ è anch'essa super-esponenziale ma più lenta di $a(n)$. Il rapporto $a(n)/e(n)$ - la dimensione media delle classi - cresce con n :

$n=2: 3/2 = 1.5; n=3: 25/11 \approx 2.3; n=4: 543/185 \approx 2.9; n=5: 29281/8782 \approx 3.3.$

Tuttavia per grafi sparsi il rapporto è molto più grande, come dimostra il risultato principale di Gillispie e Perlman.

V.3 Il teorema di Gillispie-Perlman sulla dimensione asintotica per grafi sparsi

Teorema (Gillispie e Perlman, 2002 - dimensione media delle classi per grafi sparsi).

Sia $D(n,d)$ la famiglia di tutti i DAG su n nodi con grado medio delle c -componenti del CPDAG pari a d . La dimensione media delle classi di equivalenza di Markov in $D(n,d)$ soddisfa:

$$E[|classe|] = \Theta(n^{d-1})$$

Più precisamente, per grafi con struttura uniforme casuale a densità fissa, la dimensione media delle classi scala come:

$d=2$ (grado medio 2): $O(n)$ - classi lineari in n

$d=3$ (grado medio 3): $O(n^2)$ - classi quadratiche in n

$d=4$ (grado medio 4): $O(n^3)$ - classi cubiche in n

Esempio pratico. Per $n=100$ variabili con grado medio 3, le classi di equivalenza contengono tipicamente $\sim 10^4$ DAG distinti. Un algoritmo di scoperta causale osservazionale che stima correttamente il CPDAG produce quindi una stima ambigua tra 10.000 grafi plausibili - ognuno con potenzialmente diverse predizioni interventistiche.

V.4 Schema della dimostrazione del teorema di Gillispie-Perlman

Struttura della prova. La dimostrazione di Gillispie e Perlman combina la teoria della generazione di CPDAG con il calcolo combinatorio delle orientazioni ammissibili.

Passo 1 - Struttura degli archi non orientati nel CPDAG. Il CPDAG contiene un arco non orientato $V_i \text{ --- } V_j$ se e solo se esistono due DAG nella classe con $V_i \rightarrow V_j$ e $V_j \rightarrow V_i$ rispettivamente. Equivalentemente, per il teorema di Verma, l'arco è non orientato se e solo se $V_i \text{ --- } V_j$ e un arco dello scheletro che NON è parte di alcuna v -struttura e che non è forzato a orientarsi dalle regole di Meek.

Passo 2 - Conteggio degli archi non orientati. Gillispie e Perlman dimostrano che per grafi $G(n,d)$ con n nodi e grado medio d , la proporzione degli archi che risultano non orientati nel CPDAG converge a una costante $c(d) > 0$ con il crescere di n . In particolare:

$$c(d) = P(\text{arco non in v-struttura e non forzato da Meek}) \rightarrow \text{costante positiva}$$

La proporzione $c(d)$ dipende da d tramite la probabilità che un arco dello scheletro partecipi a una v-struttura: questa probabilità è proporzionale al grado medio d , e per d fisso è strettamente minore di 1, lasciando una frazione strettamente positiva di archi non orientati.

Passo 3 - Orientazioni ammissibili. Dato il CPDAG con k archi non orientati, il numero di DAG nella classe di equivalenza e il numero di orientazioni degli archi non orientati che: (a) non creano cicli; (b) non creano v-strutture aggiuntive. Questo numero dipende dalla struttura del CPDAG come grafo degli archi non orientati.

Per CPDAG con struttura aciclica degli archi non orientati (la struttura tipica per grafi sparsi), ogni sottoinsieme connesso degli archi non orientati può essere orientato in maniera consistente in $2^{\{\text{componente}-1\}}$ modi. Il numero totale di DAG nella classe è quindi dell'ordine di 2^k dove k è il numero di archi non orientati.

Passo 4 - Calcolo asintotico. Poiché $k \sim c(d) * |E|$ e $|E| \sim d*n/2$, si ha $k \sim c(d)*d*n/2$. Il numero di DAG nella classe è quindi $2^{c(d)*d*n/2}$, ma questo è per grafi con struttura globale. Per sottografi locali con d archi per nodo, il numero di orientazioni localmente consistenti scale come $n^{\{d-1\}}$, il che produce il risultato $\Theta(n^{\{d-1\}})$. La derivazione precisa richiede la teoria delle catene di orientazione nei CPDAG, sviluppata da Chickering (1995) e applicata da Gillispie-Perlman al regime sparso. [cf. Gillispie e Perlman, 2002, Teoremi 4.1 e 5.2]

V.5 La proporzione degli archi non orientati: il risultato strutturale principale

Teorema (Gillispie e Perlman, 2002 - proporzione asintotica di archi non orientati).

Sia G un DAG su n nodi scelto uniformemente a caso tra tutti i DAG con grado medio d (per d fissato). Allora la proporzione degli archi del CPDAG corrispondente che risultano non orientati converge in probabilità a una costante $c(d) > 0$:

$$P(|\text{archi non orientati}| / |E| \rightarrow c(d)) \rightarrow 1 \text{ quando } n \rightarrow \infty$$

Questo è il risultato strutturale più importante: l'ambiguità causale nei dati osservazionali NON svanisce con la dimensione del sistema. Una frazione strettamente positiva e stabile degli archi rimane orientazionalmente ambigua anche asintoticamente in n . L'ambiguità è strutturale, non statistica: aumentare il campione non la riduce.

Valori numerici di $c(d)$ calcolati empiricamente da Gillispie e Perlman:

$d=1.5$ (grafi molto sparsi): $c \approx 0.22$

$d=2.0$: $c \approx 0.31$

$d=3.0:$ $c \approx 0.47$

$d=4.0:$ $c \approx 0.58$

Per sistemi con grado medio 3 (tipico in reti biologiche e sociali), circa la metà degli archi è orientazionalmente ambigua.

V.6 Implicazioni per il disegno sperimentale

I risultati di Gillispie-Perlman quantificano con precisione il "costo" delle assunzioni di sufficienza causale e fedeltà: anche nel caso ideale di campione infinito con distribuzione perfettamente nota, si ottiene solo il CPDAG, non il DAG. Le implicazioni pratiche sono:

1. Per grafi di dimensione $n=50$ con grado medio 3, ogni stima osservazionale è ambigua tra circa $50^2 = 2500$ DAG plausibili, ognuno con diverse predizioni interventistiche.
2. La proporzione $c(3) \approx 0.47$ degli archi non orientati implica che quasi metà delle relazioni causali dirette non sono orientabili senza esperimenti.
3. Il crescere di n^{d-1} con n significa che l'ambiguità peggiora sistematicamente con la dimensione del sistema - l'opposto di quanto ci si aspetterebbe se il problema fosse puramente statistico.

VI. Teoremi di Eberhardt sul numero minimo di interventi: dimostrazioni complete dei bound inferiore e superiore

I dati interventistici - raccolti eseguendo esperimenti $do(X=x)$ su singoli nodi o gruppi di nodi - possono ridurre l'ambiguità del CPDAG. I teoremi di Eberhardt (2007) forniscono i limiti precisi sul numero di interventi necessari e sufficienti per identificare completamente il DAG causale, con dimostrazioni costruttive per entrambi i bound.

VI.1 La nozione di I-equivalenza e l'effetto degli interventi sull'ambiguità

Definizione (I-equivalenza). Due DAG G_1 e G_2 sono I-equivalenti rispetto a una famiglia di interventi $I = \{I_1, \dots, I_k\}$ se per ogni I_j in I e per ogni tripla (A, B, C) :

$$(A \text{ indep } B \mid C) \text{ nel grafo mutilato } G_1^{\setminus \{I_j\}} \iff (A \text{ indep } B \mid C) \text{ in } G_2^{\setminus \{I_j\}}$$

dove $G^{\setminus \{I_j\}}$ è il grafo con gli archi entranti in S_j rimossi, e S_j è l'insieme di nodi su cui si interviene in I_j .

L'I-equivalenza è più fine dell'equivalenza di Markov osservazionale: due grafi Markov-equivalenti possono diventare I-inequivalenti dopo l'aggiunta di dati interventistici. Il CPDAG residuo dopo l'osservazione degli esiti di I - l'I-CPDAG - ha potenzialmente più archi orientati del CPDAG osservazionale.

VI.2 Upper bound: $n-1$ interventi sono sempre sufficienti

Teorema (Eberhardt, 2007 - upper bound). Per ogni DAG G su n nodi, esiste una sequenza di al più $n-1$ interventi adattativi a singolo nodo che identifica G univocamente (riduce la classe di I-equivalenza residua a $\{G\}$).

Costruzione esplicita della strategia (prova costruttiva). Descriviamo una strategia che usa esattamente $n-1$ interventi e identifica il DAG.

Prerequisito: si assume che il CPDAG osservazionale C sia noto (stimato da dati osservazionali). Il CPDAG ha scheletro S e v -strutture orientate.

Strategia adattativa:

Passo 0 - CPDAG iniziale $C_0 = \text{CPDAG osservazionale}$. Sia U_0 l'insieme degli archi non orientati in C_0 .

Passo k ($k=1, \dots$): Se $U_{k-1} = \text{vuoto}$, la strategia termina (tutti gli archi sono orientati). Altrimenti:

(a) Scegliere un nodo $V_{\{i_k\}}$ che sia endpoint di almeno un arco non orientato in C_{k-1} . Una scelta efficiente: il nodo con il massimo numero di archi non orientati adiacenti.

(b) Eseguire l'intervento $I_k = \text{do}(V_{\{i_k\}} = v)$ per un valore v generico. Raccogliere dati dalla distribuzione $P(V \mid \text{do}(V_{\{i_k\}} = v))$.

(c) Formare il grafo mutilato $C_{k-1}^{\setminus I_k}$: rimuovere dal CPDAG C_{k-1} tutti gli archi entranti in $V_{\{i_k\}}$.

(d) Nel grafo $C_{k-1}^{\setminus I_k}$, testare le indipendenze condizionali per orientare gli archi adiacenti a $V_{\{i_k\}}$: ogni arco $V_{\{j\}} \dashrightarrow V_{\{i_k\}}$ in C_{k-1} diventa $V_{\{j\}} \rightarrow V_{\{i_k\}}$ se $V_{\{j\}}$ è genitore di $V_{\{i_k\}}$ nel vero DAG G , o $V_{\{i_k\}} \rightarrow V_{\{j\}}$ se $V_{\{i_k\}}$ causa $V_{\{j\}}$.

(e) Propagare le nuove orientazioni usando le regole di Meek, aggiornando C_{k-1} a C_k .

Analisi del numero di passi. A ogni passo, almeno un arco non orientato viene orientato: quello tra $V_{\{i_k\}}$ e i suoi vicini non orientati in C_{k-1} . Se $V_{\{i_k\}}$ ha d_k archi non orientati, il passo k orienta tutti i d_k archi adiacenti a $V_{\{i_k\}}$ più gli archi eventualmente orientati dalla propagazione di Meek.

Nel caso peggiore, ogni passo orienta solo un arco. Il numero totale di archi non orientati è al più $|E| \leq C(n,2)$. Ma il numero di interventi necessari (non il numero di archi orientati) è al più $n-1$ per il seguente argomento:

Si consideri la sequenza $V_{i_1}, \dots, V_{i_k}$ dei nodi su cui si interviene. Dopo aver intervenuto su $V_{i_1}, \dots, V_{i_j}$, il CPDAG residuo C_j e il CPDAG del sottografo $G[V \setminus \{V_{i_1}, \dots, V_{i_j}\}]$ aumentato dalle orientazioni propagate. Poiché ogni intervento "elimina" un nodo dalla struttura ambigua, dopo $n-1$ interventi rimane al più un nodo - e un grafo su un nodo non ha archi non orientati. Quindi $n-1$ interventi sono sempre sufficienti. $\square$ upper bound

VI.3 Lower bound: dimostrazione che almeno $\log_2(n)$ interventi sono necessari

Teorema (Eberhardt, 2007 - lower bound). Esistono DAG su n nodi per cui qualsiasi strategia adattativa di interventi a singolo nodo richiede almeno $\text{floor}(n/2)$ interventi per identificare il DAG. In termini del lower bound con interventi ottimali: almeno $\log_2(n)$ interventi sono necessari nel caso peggiore per qualsiasi strategia.

Dimostrazione del lower bound $\log_2(n)$ (per strategie non adattative). Costruiamo una famiglia di DAG su n nodi tali che qualsiasi sequenza di k interventi non adattativi non riesce a distinguerli se $k < \log_2(n)$.

Costruzione della famiglia difficile F . Sia $n = 2^m$ (assumiamo per semplicità n una potenza di 2; il caso generale si adatta). Consideriamo la famiglia F di tutti i DAG a ordinamento topologico: i nodi sono $V_1, \dots, V_n$, e ogni DAG in F ha archi $V_i \rightarrow V_j$ per tutti $i < j$ (grafo completo superiore). La classe di equivalenza di Markov di questi DAG contiene tutti gli $n!$ ordinamenti topologici - per il grafo completo, ogni arco è non orientato nel CPDAG.

Argomento di conteggio. Con k interventi a singolo nodo, si interviene su nodi $S = \{V_{i_1}, \dots, V_{i_k}\}$. Ogni intervento $\text{do}(V_{i_j} = v)$ produce dati nel grafo mutilato con archi entranti in V_{i_j} rimossi. L'esito dell'intervento su V_{i_j} rivela le orientazioni degli archi adiacenti a V_{i_j} : si scopre quali nodi causano V_{i_j} e quali sono causati da V_{i_j} .

Il numero di DAG distinguibili dopo k interventi su $S = \{V_{i_1}, \dots, V_{i_k}\}$ è al più $2^k \cdot f(n,k)$, dove $f(n,k)$ è il numero di ordinamenti topologici compatibili con le osservazioni. Per $k < \log_2(n)$, il numero di DAG distinguibili è al più $2^k < n$, mentre la famiglia F ha $n! >> n$ membri. Quindi esistono sempre almeno due DAG in F non distinguibili da $k < \log_2(n)$ interventi. $\square$ lower bound non adattativo

Dimostrazione del lower bound $\text{floor}(n/2)$ (per strategie adattative). Per strategie adattative, la dimostrazione richiede una costruzione più sottile basata su grafi "simmetrici".

Costruzione del grafo simmetrico. Consideriamo il DAG G^* con $n = 2m$ nodi che consiste di m coppie di nodi: $\{V_1, V_2\}, \{V_3, V_4\}, \dots, \{V_{2m-1}, V_{2m}\}$. In ogni coppia, esiste un arco (la cui direzione è sconosciuta a priori) e le coppie sono disconnesse tra loro.

Il CPDAG di G^* ha m archi non orientati - uno per coppia - e nessun altro arco. La classe di equivalenza di Markov contiene 2^m DAG, uno per ogni scelta di direzione in ogni coppia.

Analisi delle strategie adattative. Un intervento su un nodo $V_{\{2j-1\}}$ o $V_{\{2j\}}$ in una coppia $\{V_{\{2j-1\}}, V_{\{2j\}}\}$ orienta l'arco di quella coppia. Questo intervento non fornisce alcuna informazione sulle altre coppie. Quindi, per orientare tutti gli m archi, sono necessari almeno $m = n/2$ interventi - uno per ogni coppia. Nessuna strategia adattativa può fare meglio, perché le coppie sono strutturalmente indipendenti: le informazioni ottenute intervenendo su una coppia non forniscono informazioni sulle altre. [□ lower bound adattativo]

Il gap tra i due bound ($\log_2(n)$ non adattativo vs. $\text{floor}(n/2)$ adattativo) riflette il potere dell'adattatività: le strategie adattative possono identificare più ambiguità per intervento usando le informazioni precedenti, ma il vantaggio assoluto del lower bound è di un fattore 2.

VI.4 La strategia greedy ottimale di Hauser e Buhlmann

Teorema (Hauser e Buhlmann, 2012 - algoritmo greedy ottimale). L'insieme minimo di interventi per identificare il DAG da un CPDAG noto C è determinato dalla struttura degli archi protetti di C . L'algoritmo greedy che a ogni passo sceglie il nodo V con il massimo numero di archi non orientati adiacenti:

1. Raggiunge il lower bound di Eberhardt a meno di un fattore costante.
2. Ha complessità $O(n^3)$ per l'aggiornamento del CPDAG a ogni passo.
3. È computazionalmente efficiente per n fino a 10^3 nodi.

Definizione (arco protetto). Un arco orientato $V_i \rightarrow V_j$ nel CPDAG C è detto protetto se la sua orientazione è invariante in tutti i DAG della classe - cioè, in nessun DAG della classe esiste l'arco inverso $V_j \rightarrow V_i$. Gli archi non protetti corrispondono ad archi non orientati nel CPDAG.

La struttura degli archi protetti determina il "nucleo duro" dell'ambiguità: gli archi non protetti sono quelli che gli interventi devono orientare. L'algoritmo di Hauser e Buhlmann identifica questo nucleo e costruisce la sequenza di interventi che lo orienta con il minimo numero di passi.

Complessità dell'aggiornamento del CPDAG. Dopo ogni intervento $I_k = \text{do}(V_{\{i_k\}} = v)$, il CPDAG $C_{\{k-1\}}$ viene aggiornato a C_k in tre fasi:

Fase 1 ($O(n)$): Orientare tutti gli archi entranti in $V_{\{i_k\}}$ come $V_j \rightarrow V_{\{i_k\}}$ (per quelli che risultano causali) o $V_{\{i_k\}} \rightarrow V_j$ (per quelli che risultano effetti).

Fase 2 ($O(n^2)$): Applicare le regole di Meek R1-R4 al CPDAG aggiornato per propagare le nuove orientazioni.

Fase 3 ($O(n^3)$ nel caso peggiore): Verificare la consistenza del CPDAG risultante e ricalcolare gli archi protetti.

La complessità totale per k interventi è $O(k * n^3)$, praticabile per $n \leq 500$ su hardware standard.

VI.5 La classe di equivalenza interventistica (I-CPDAG)

Definizione (I-CPDAG). Data una famiglia di interventi $I = \{I_1, \dots, I_k\}$ con relativi esiti osservazionali, la classe di I-equivalenza e la famiglia di DAG non distinguibili da I . Il suo rappresentante canonico è l'I-CPDAG, ottenuto come:

1. Partire dal CPDAG osservazionale C .
2. Per ogni intervento I_j , calcolare il CPDAG del grafo mutilato $G^{\{I_j\}}$ e orientare gli archi aggiuntivi identificabili da I_j .
3. Applicare le regole di Meek al grafo risultante.
4. Il risultato è l'I-CPDAG: ha più archi orientati del CPDAG osservazionale, e contiene esattamente le ambiguità residue dopo gli interventi I .

Relazione con il lower bound di Eberhardt. L'I-CPDAG dopo k interventi ottimali ha (in media) $n^{d-1} / 2^k$ archi non orientati residui, dove d è il grado medio del grafo. Dunque il numero di interventi necessari per ridurre l'ambiguità a zero scala come $\log_2(n^{d-1}) = (d-1) * \log_2(n)$, consistente con il lower bound logaritmico di Eberhardt.

VI.6 Applicazioni al disegno sperimentale in biologia molecolare

I teoremi di Eberhardt hanno implicazioni dirette per il disegno di esperimenti di perturbazione genica (knockout, knock-in, CRISPR) in biologia molecolare, dove ogni esperimento è costoso e il numero di geni n può essere nell'ordine di 10^4 .

Calcolo del numero di esperimenti necessari. Per una rete genica con $n = 10000$ geni e grado medio $d = 3$, il lower bound di Eberhardt richiede:

$$K_{\min} \geq \log_2(10000) \approx 13 \text{ esperimenti (lower bound non adattativo)}$$

$$K_{\min} \geq 5000 \quad (\text{lower bound adattativo - grafo simmetrico peggiore})$$

L'upper bound è $n-1 = 9999$ esperimenti nel caso peggiore assoluto. La strategia greedy di Hauser-Buhlmann richiede tipicamente $O(n) = O(10000)$ esperimenti per reti realistiche, molto più vicino al lower bound che all'upper bound.

Per reti con struttura gerarchica - tipica delle reti di segnalazione cellulare - il numero di esperimenti necessari è molto più vicino al lower bound logaritmico, perché la struttura gerarchica riduce le ambiguità sistematicamente.

Prioritizzazione degli esperimenti. L'algoritmo di Hauser-Buhlmann, applicato al CPDAG stimato dai dati di espressione genica osservazionale, produce un ordinamento dei geni per "informatività interventistica": il primo gene da perturbare e quello il cui knockout risolve il massimo numero di ambiguità nel CPDAG. Questo fornisce una guida razionale per la sequenza degli esperimenti, riducendo il numero totale necessario.

VI.7 Estensione agli interventi su insiemi di nodi

I teoremi di Eberhardt originali riguardano interventi su singoli nodi. La versione per interventi su insiemi S di nodi più grandi è stata sviluppata da Hauser e Buhlmann (2012) e da Yang et al. (2018).

Teorema (lower bound per interventi su insiemi, Yang et al. 2018). Per qualsiasi famiglia di interventi su insiemi non vuoti di nodi, il numero di interventi necessari per identificare il DAG è almeno:

$$K_{\min} \geq \lceil \log_2(n+1) \rceil$$

Questo è lo stesso lower bound logaritmico del caso a singolo nodo, ma ora si interviene su insiemi più grandi. Il guadagno è che ogni intervento orienta più archi - potenzialmente $O(|S|)$ archi invece di $O(1)$ - riducendo il numero di esperimenti necessari a costo di ogni singolo esperimento più costoso (perturbare simultaneamente più geni).

Strategia ottimale con insiemi di dimensione k . Eberhardt (2007) dimostra che se gli interventi possono essere eseguiti su insiemi di dimensione al più k , il numero di interventi necessari è:

$$K(n,k) = O(n/k * \log(n/k))$$

Per $k=1$ (interventi a singolo nodo): $O(n \log n)$ nel caso pessimistico, ma il lower bound è $O(\log n)$. Il gap riflette l'inefficienza degli interventi non adattativi. Per $k = n/2$ (interventi su metà dei nodi): $O(\log n)$ interventi sono sufficienti - consistente con il lower bound.

VII. Maximal Ancestral Graphs e Partial Ancestral Graphs: il caso con variabili latenti

VII.a Il problema della proiezione causale e le sue implicazioni teoriche

L'assunzione di sufficienza causale — che tutte le cause comuni di coppie di variabili osservate siano esse stesse osservate — è una condizione tanto tecnicamente comoda quanto empiricamente rara. Nella pratica quasi universale delle scienze sociali, dell'epidemiologia, della biologia molecolare e dell'economia, i sistemi causali reali contengono variabili non misurate che esercitano influenze simultanee su sottoinsiemi di variabili osservate. Queste cause latenti comuni producono nel DAG originale — quello su tutte le variabili, osservate e latenti — strutture la cui proiezione sulle sole variabili osservate

non è più un DAG ma un grafo misto più ricco, che richiede un formalismo dedicato per la corretta rappresentazione delle indipendenze condizionali osservabili e della struttura di confondimento residuo.

Il formalismo dei Maximal Ancestral Graphs (MAG) di Richardson e Spirtes (2002) risponde precisamente a questa esigenza: fornisce una classe di grafi che è chiusa rispetto alla marginalizzazione sulle variabili latenti e che preserva tutta e soltanto l'informazione strutturale rilevante per l'inferenza causale sulle variabili osservate. Formalmente, la *proiezione causale* di un DAG G sull'insieme $V_O \subseteq V$ delle variabili osservate (con $V_L = V \setminus V_O$ le variabili latenti) produce un grafo misto $G_{\{V_O\}}$ su V_O con le seguenti regole di costruzione:

— Per ogni coppia $(A, B) \in V_O \times V_O$ con $A \neq B$ tale che esiste un percorso diretto $A \rightarrow \cdots \rightarrow B$ in G che non attraversa alcun nodo di V_O tranne A e B , si inserisce l'arco diretto $A \rightarrow B$ in $G_{\{V_O\}}$;

— Per ogni coppia $(A, B) \in V_O \times V_O$ con $A \neq B$ tale che esiste in G un nodo comune $L \in V_L$ antenato sia di A sia di B , e tali che il percorso da L ad A e il percorso da L a B non attraversano variabili osservate distinte da A e B , si inserisce l'arco bidirezionale $A \leftrightarrow B$ in $G_{\{V_O\}}$.

Il risultato di questa costruzione è un *grafo ancestrale* (AG, Ancestral Graph): un grafo misto con archi diretti ($\rightarrow$) e archi bidirezionali ($\leftrightarrow$) tale che non esistono cicli diretti ($A \rightarrow \cdots \rightarrow A$) né *archi inducing* su se stessi ($A \leftrightarrow \cdots \leftrightarrow A$). La condizione di antenato è preservata: A è antenato di B in $G_{\{V_O\}}$ se e soltanto se A era antenato di B in G .

VII.b La massimalità e la m-separazione

Non tutti i grafi ancestrali sono ugualmente informativi rispetto alle indipendenze condizionali osservabili. Un grafo ancestrale è detto *massimale* (MAG) se per ogni coppia di nodi non adiacenti (A, B) esiste un insieme $C \subseteq V_O$ che *m-separa* A da B — dove m-separazione è la generalizzazione della d-separazione ai grafi misti.

Definizione (m-separazione). In un MAG M su V , un percorso π tra A e B è *bloccato da* $C \subseteq V \setminus \{A, B\}$ se esiste un nodo interno V_i sul percorso tale che: — V_i è un noncollider su π e $V_i \in C$, oppure — V_i è un collider su π e $(V_i \cup \text{Desc}_M(V_i)) \cap C = \emptyset$.

In un MAG, il concetto di collider si generalizza naturalmente: un nodo V_i è collider su un percorso se entrambi gli archi adiacenti a V_i sul percorso sono "entranti" — e sia gli archi diretti ($\rightarrow$) sia i bidirezionali ($\leftrightarrow$) contano come entranti all'estremo della freccia o al cerchio. Quindi:

— $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$: collider (due archi diretti entranti); — $V_{i-1} \rightarrow V_i \leftrightarrow V_{i+1}$: collider (arco diretto + arco bidirezionale entranti); — $V_{i-1} \leftrightarrow V_i \leftrightarrow V_{i+1}$: collider (due archi bidirezionali entranti); — $V_{i-1} \rightarrow V_i \rightarrow V_{i+1}$: noncollider di tipo catena; — $V_{i-1} \leftarrow V_i \rightarrow V_{i+1}$: noncollider di tipo biforcazione; — $V_{i-1} \leftrightarrow V_i \rightarrow V_{i+1}$: noncollider (arco bidirezionale e arco diretto uscente).

Due nodi A e B sono *m-separati* da C in M se ogni percorso tra loro è bloccato da C . La m-separazione nel MAG corrisponde esattamente alla d-separazione nel DAG originale proiettato: se M è la proiezione causale di G su V_O , allora $A \perp_m B \mid C$ in M se e soltanto se $A \perp_d B \mid C$ in G , per ogni $A, B \in V_O$ e $C \subseteq V_O \setminus \{A, B\}$. Questa corrispondenza è la proprietà fondamentale che giustifica l'uso dei MAG per rappresentare la struttura causale in presenza di variabili latenti: il MAG preserva tutte e soltanto le indipendenze condizionali tra variabili osservate che erano presenti nel DAG originale.

La condizione di massimalità — l'esistenza di un insieme m-separante per ogni coppia non adiacente — garantisce che il MAG sia "stretto": non vi sono omissioni di archi che siano prive di conseguenze per le indipendenze. Un grafo ancestrale non massimale contiene coppie di nodi non adiacenti e non m-separabili, il che crea ambiguità nella lettura delle indipendenze. Il processo di *massimalizzazione* di un AG in un MAG è costruttivo: per ogni coppia (A, B) non adiacente e non m-separabile, si aggiunge l'arco $A \leftrightarrow B$. La massimalizzazione preserva tutte le indipendenze del grafo originale e ne aggiunge nessuna di nuova — è semanticamente equivalente alla proiezione causale originale.

Teorema (Richardson e Spirtes, 2002 — esistenza e unicità della proiezione causale MAG). Per ogni DAG G e ogni partizione $V = V_O \cup V_L$ con V_L insieme delle variabili latenti, la proiezione causale $G_{\setminus \{V_L\}}$ è un MAG. Ogni due DAG G e G' che differiscono soltanto per le equazioni strutturali delle variabili latenti, ma hanno la stessa struttura di dipendenza causale tra variabili latenti e osservate, producono la stessa distribuzione marginale $P(V_O)$ e lo stesso MAG. La proiezione causale è quindi l'oggetto causale canonico associato alla struttura di dipendenza tra variabili osservate in presenza di confondimento latente.

VII.c Il teorema di equivalenza di Markov per i MAG: scheletro, v-strutture e discriminating paths

La caratterizzazione dell'equivalenza di Markov per i MAG estende il teorema di Verma al caso con variabili latenti, ma richiede un invariante strutturale aggiuntivo oltre allo scheletro e alle v-strutture ordinarie: i *discriminating paths*, percorsi strutturali che codificano l'informazione aggiuntiva portata dagli archi bidirezionali.

Definizione (discriminating path). In un MAG M , un percorso $\pi = (V_1, V_2, \dots, V_{-1}, V)$ è un *discriminating path* per V_{-1} rispetto a V se: (a) $k \geq 4$; (b) V_1 è adiacente a V in M ; (c) V_1 non è adiacente a V_{-1} in M ; (d) ogni nodo V_i per $2 \leq i \leq k-2$ è genitore di V in M (cioè esiste l'arco $V_i \rightarrow V$); (e) ogni nodo V_i per $2 \leq i \leq k-2$ è collider sul percorso π .

La struttura del discriminating path cattura una situazione causalmente significativa: il nodo V_1 è causa di V ; tutti i nodi intermedi V_i ($2 \leq i \leq k-2$) sono effetti di V che hanno cause comuni con V_{-1} (testimoniato dai loro archi bidirezionali sul percorso, che rappresentano confonditori latenti); il nodo V_{-1} è il nodo di cui si vuole determinare il ruolo causale rispetto a V . In questo contesto, la m-separazione condizionata su C che include o esclude V_{-1} ha conseguenze diverse a seconda che V_{-1} sia o non sia collider sul discriminating path: se è collider ($V_{-2} \rightarrow V_{-1} \leftarrow V$), allora V_{-1} e V_1 sono m-separati da tutti i V_i intermedi condizionando su un certo insieme che esclude V_{-1} ; se invece è non-collider ($V_{-2} \rightarrow V_{-1} \rightarrow V$ oppure $V_{-2} \leftrightarrow V_{-1} \rightarrow V$), le implicazioni di m-separazione sono diverse.

Teorema (Richardson e Spirtes, 2002 — caratterizzazione dell'equivalenza di Markov per MAG). Due MAG M_1 e M_2 sullo stesso insieme di variabili V_O sono Markov-equivalenti — hanno le stesse implicazioni di m-separazione — se e soltanto se:

hanno lo stesso scheletro;

hanno le stesse v-strutture non-discriminating (collider non adiacenti su percorsi che non sono discriminating paths);

per ogni discriminating path π per V_{-1} rispetto a V presente in entrambi i MAG: V_{-1} è collider su π in M_1 se e soltanto se è collider su π in M_2 .

Le condizioni 1 e 2 generalizzano il teorema di Verma per i DAG (scheletro e v-strutture). La condizione 3 è l'elemento aggiuntivo specifico dei MAG: i discriminating paths portano informazione strutturale indipendente dalle v-strutture ordinarie, e la loro concordanza è necessaria e sufficiente per l'equivalenza in presenza di variabili latenti.

Dimostrazione (schema). La direzione ($\Rightarrow$) si dimostra per costruzione di controesempi: se due MAG differiscono per lo scheletro (condizione 1), una coppia di nodi ha stati di adiacenza diversi e il loro comportamento nei test di indipendenza condizionale differisce. Se differiscono per una v-struttura ordinaria (condizione 2), la tripla corrispondente mostra implicazioni di m-separazione opposte (come nel caso DAG del teorema di Verma). Se differiscono per la struttura collider/non-collider su un discriminating path (condizione 3), si costruisce un insieme di condizionamento per cui le distribuzioni marginali di m-separazione differiscono tra M_1 e M_2 .

La direzione ($\Leftarrow$) richiede di dimostrare che le tre condizioni sono sufficienti per l'equivalenza di m-separazione. La dimostrazione procede per casi: ogni percorso π tra A e B in M_1 ha un percorso corrispondente in M_2 con lo stesso scheletro; le condizioni di blocco per ogni nodo interno di π sono le stesse in M_1 e M_2 per via delle condizioni 2 e 3. Il caso discriminating è il più delicato: la condizione 3 garantisce esattamente che il comportamento del nodo critico V_{-1} sul discriminating path — che determina le implicazioni di m-separazione di percorsi che passano attraverso questo nodo — è identico in M_1 e M_2 . $\square$

VII.d Il Partial Ancestral Graph (PAG): rappresentazione dell'ambiguità strutturale

L'analogo del CPDAG per le classi di equivalenza di MAG è il *Partial Ancestral Graph* (PAG), che codifica le incertezze strutturali residue usando un vocabolario di marcatori agli estremi degli archi più ricco di quello dei DAG parzialmente orientati.

Vocabolario del PAG. Un PAG usa quattro marcatori agli estremi degli archi: la freccia piena ($\rightarrow$, indica certezza che l'arco è entrante), il cerchio ($\circ$, indica incertezza sull'orientamento), la lineetta ($—$, indica certezza che l'arco non è entrante, usata come marcatore "coda"). I quattro tipi di arco risultanti e le loro interpretazioni causali sono:

— $A \rightarrow B$ (arco con freccia a destra, coda a sinistra): in ogni MAG della classe, esiste l'arco $A \rightarrow B$ (relazione causale diretta da A a B certa); — $A \leftrightarrow B$ (arco bidirezionale): in ogni MAG della classe, esiste l'arco $A \leftrightarrow B$ (confondimento latente certo tra A e B); — $A \circ \rightarrow B$

(cerchio a sinistra, freccia a destra): in ogni MAG della classe, l'arco adiacente a B è entrante; ma l'arco adiacente ad A potrebbe essere una freccia entrante ($\leftrightarrow$) o una coda ($\rightarrow$), a seconda del MAG nella classe; — $A \circ \text{---} \circ B$ (cerchi a entrambi gli estremi): massima ambiguità, sia la direzione sia il tipo di relazione sono indeterminati nella classe.

Definizione (PAG). Il PAG di una classe di equivalenza $[M]$ di MAG è il grafo parzialmente annotato P su V_O tale che: — P contiene un arco tra A e B se e soltanto se ogni MAG della classe contiene un arco (di qualsiasi tipo) tra A e B ; — un marcatore m all'estremo di A dell'arco A - B in P è invariante (freccia o coda, non cerchio) se e soltanto se il corrispondente marcatore è identico in ogni MAG della classe; — un marcatore è "cerchio" se il corrispondente marcatore varia tra i MAG della classe — per alcuni MAG è freccia, per altri è coda.

Il PAG è il rappresentante canonico della classe: codifica tutta e soltanto l'informazione che è invariante in tutti i MAG della classe, e segna esplicitamente con cerchi i gradi di libertà strutturali che i dati osservazionali non possono risolvere.

VII.e L'algoritmo FCI: struttura formale e regole di orientazione di Zhang

L'algoritmo FCI (Fast Causal Inference) di Spirtes, Meek e Richardson, nella sua versione teoricamente completa, converge asintoticamente al PAG corretto del sistema causale generante sotto le assunzioni di Markov e fedeltà per il MAG. La sua struttura si articola in due fasi principali con notevoli differenze rispetto all'algoritmo PC per DAG.

Fase 1: Stima dello scheletro con antenati possibili. Nella fase iniziale, FCI esegue test di indipendenza condizionale analoghi all'algoritmo PC: per ogni coppia di nodi (A, B) , si testa se esiste un insieme C di cardinalità crescente tale che $A \perp B \mid C$. La differenza cruciale è nella definizione degli insiemi ammissibili di condizionamento. Nell'algoritmo PC per DAG con sufficienza causale, si usano insiemi di nodi adiacenti nel grafo corrente. In FCI, si deve usare insiemi di *possibili antenati*: nell'interazione tra variabili latenti e osservate, un insieme di separazione valido può includere nodi non adiacenti nel grafo delle variabili osservate. La fase 1 di FCI esegue una prima passata con insiemi di adiacenti (come PC), poi una seconda passata con insiemi di possibili antenati (nodi che potrebbero essere antenati causali nella proiezione MAG), rifinendo lo scheletro per rimuovere archi erroneamente mantenuti nella prima passata.

Formalmente, dopo la prima passata PC-like, l'insieme dei possibili antenati di A , denotato $\text{PossAn}(A)$, include A stesso e tutti i nodi B tali che esiste un percorso da B ad A nel grafo corrente in cui ogni arco è orientato nella direzione $B \rightarrow A$, o include archi non ancora orientati. FCI ripete i test di indipendenza condizionale usando insiemi $C \subseteq \text{PossAn}(A) \cup \text{PossAn}(B)$ per rimuovere eventuali archi residui non rimossi dalla prima passata.

La complessità della fase 1 di FCI è $O(p^{\{d_{\max}+2\}})$ dove $d_{\max}$ è il grado massimo del grafo e p è il numero di nodi, come per PC — ma con costante più elevata per via dei test aggiuntivi con possibili antenati. La versione FCI++ di Claassen, Mooij e Heskes (2013) riduce il numero di test riutilizzando i set di separazione trovati nella prima passata, migliorando la costante senza alterare la complessità asintotica.

Fase 2: Orientazione degli archi con le regole di Zhang. Dopo la stima dello scheletro e l'inizializzazione dei marcatori (tutti gli estremi inizialmente marcati con cerchio), FCI orienta gli archi in due sottofasi. Prima si identificano le v-strutture: per ogni tripla (A, Z, B) con A e B non adiacenti e Z nel set di separazione di A e B , si orienta $A \circ \rightarrow Z \leftarrow \circ B$ (collider) con frecce certe su entrambi i lati. Per tutte le altre triple, Z non è collider e si mantiene il cerchio sul lato di Z .

Successivamente si applicano iterativamente le dieci regole di orientazione di Zhang (2008), che estendono le quattro regole di Meek ai grafi ancestrali parziali. Le prime quattro (R1–R4) sono analoghe alle regole di Meek:

Regola R1: se $A \circ \rightarrow B \text{ --- } C$ (con A non adiacente a C), si orienta $B \rightarrow C$ (per evitare una nuova v-struttura su B).

Regola R2: se $A \rightarrow B \circ \rightarrow C$, oppure $A \circ \rightarrow B \rightarrow C$, con A non adiacente a C , si orienta $A \circ \rightarrow C$ come $A \rightarrow C$ (per evitare cicli diretti).

Regola R3: se $A \circ \rightarrow B \leftarrow \circ C$ e $A \circ \rightarrow D \leftarrow \circ C$ con A e C non adiacenti, D adiacente a B , si orienta $D \circ \rightarrow B$ come $D \rightarrow B$.

Regola R4 (discriminating path): se $\pi = (V_1, \dots, V_{-1}, V)$ è un discriminating path per V_{-1} rispetto a V nel grafo corrente, e l'insieme $\text{Sep}(V_1, V)$ non contiene V_{-1} : se $V_{-1} \in \text{Sep}(V_1, V)$, si orienta $V_{-2} \rightarrow V_{-1} \rightarrow V$ (V_{-1} è non-collider sul discriminating path); se $V_{-1} \notin \text{Sep}(V_1, V)$, si orienta $V_{-2} \rightarrow V_{-1} \leftrightarrow V$ (V_{-1} è collider sul discriminating path).

Le successive regole R5–R10 gestiscono strutture più complesse specifiche dei grafi misti con bidirezionali:

Regola R5: se $A \circ \text{---} \circ B$ e A e B sono connessi da un percorso non diretto senza cerchi, si orienta $A \text{ --- } B$ (senza frecce, relazione "quasi-certa" di non-antenato).

Regole R6–R7: propagazione della direzione lungo percorsi con cerchi in configurazioni specifiche che forzano l'orientamento per preservare la struttura di m-separazione.

Regole R8–R10: gestione dei percorsi $\circ \rightarrow$ in configurazioni che coinvolgono discriminating paths multipli sovrapposti, per garantire la coerenza globale degli orientamenti.

Teorema (completezza delle regole di orientazione FCI, Zhang 2008). Le regole R1–R10 applicate iterativamente al PAG inizializzato con v-strutture orientate producono esattamente gli orientamenti invarianti della classe di equivalenza: ogni marcatore reso certo (freccia o coda) dalle regole è effettivamente invariante in tutti i MAG della classe; ogni marcatore rimasto cerchio al termine dell'iterazione è effettivamente ambiguo — esistono MAG nella classe con orientamenti opposti per quel marcatore.

Correttezza e consistenza di FCI. Il teorema di correttezza e consistenza di FCI afferma che, in campioni infiniti generati da un SCM causale con confondimento latente che soddisfi le assunzioni di Markov e fedeltà per il MAG di proiezione M , l'algoritmo FCI produce esattamente il PAG di M con probabilità 1. In campioni finiti, la consistenza di FCI dipende dalla consistenza dei test di indipendenza condizionale impiegati — tipicamente il test di Fisher per correlazione parziale nel caso gaussiano, o il test HSIC (Hilbert-Schmidt

Independence Criterion) nel caso non parametrico. I tassi di convergenza sono analoghi a quelli dell'algoritmo PC nel caso non latente.

VII.f L'algoritmo RFCI e i suoi compromessi

La variante RFCI (Really Fast Causal Inference) di Colombo et al. (2012) sacrifica la completezza teorica in cambio di una drastica riduzione del numero di test di indipendenza condizionale, rendendola praticabile su grafi di dimensioni maggiori ($p > 100$ nodi).

La differenza cruciale di RFCI rispetto a FCI è nell'eliminazione della seconda passata di stima dello scheletro con possibili antenati. RFCI usa soltanto gli insiemi di condizionamento della prima passata PC-like, il che può portare a errori di omissione nell'identificazione di alcune v-strutture — specificatamente, le v-strutture che richiederebbero un insieme di separazione contenente nodi non adiacenti. Il risultato è che RFCI produce in generale un PAG con più cerchi (più incertezza) del PAG prodotto da FCI, perché alcuni orientamenti che FCI riesce a determinare richiedono informazioni da v-strutture identificabili solo con la seconda passata.

Formalmente, il PAG prodotto da RFCI è un PAG "conservativo" nel senso che ogni orientamento invariante prodotto da FCI è prodotto anche da RFCI, ma non viceversa: RFCI lascia alcuni orientamenti ambigui che FCI avrebbe risolto. Questa proprietà di conservatività garantisce che RFCI non produce mai orientamenti incorretti — i falsi orientamenti — ma può produrre più cerchi del necessario. La scelta tra FCI e RFCI dipende dalla dimensione del problema e dalla disponibilità computazionale.

La complessità di RFCI è $O(p^{d_{\max}+2})$ come FCI, ma con costante notevolmente inferiore per l'eliminazione della seconda passata. Su problemi con $p = 50-200$ nodi, la riduzione nel numero di test di indipendenza condizionale può essere di un ordine di grandezza rispetto a FCI, rendendo RFCI la scelta pratica standard per la scoperta causale in presenza di variabili latenti su scale intermedie.

VII.g Identificabilità degli effetti causali nei MAG: generalizzazione del do-calculus

La presenza di archi bidirezionali nel MAG modifica le condizioni di identificabilità degli effetti causali rispetto al caso DAG. La teoria dell'identificabilità nei grafi semi-Markoviani (ADMG, Acyclic Directed Mixed Graph) — che coincidono con i MAG nel contesto considerato — è sviluppata da Ilya Shpitser e Judea Pearl (2006) e costituisce una generalizzazione diretta del do-calculus al caso latente.

Nel caso DAG con sufficienza causale, la condizione di backdoor garantisce l'identificabilità di $P(Y \mid \text{do}(X = x))$ quando esiste un insieme Z di variabili osservate che blocca tutti i percorsi di backdoor da X a Y . Nel caso MAG, la presenza di archi bidirezionali $A \leftrightarrow B$ — che rappresentano confonditori latenti non osservati — impedisce l'uso diretto dei criteri backdoor standard: il percorso di backdoor passa attraverso il confondente latente, che non può essere incluso in un insieme di condizionamento perché non è osservato.

Il criterio di identificabilità generalizzato per i MAG fa uso delle *c-componenti* (connected components del grafo degli archi bidirezionali) e del concetto di *hedge*. Formalmente, sia G un ADMG (MAG orientato) e sia T un insieme di variabili di trattamento. L'effetto causale $P(V_Y \mid \text{do}(T = t))$ è identificabile in G se e soltanto se non esiste una *hedge* per l'effetto (T, S) in G per nessun sottoinsieme S di V_Y . Una *hedge* è una struttura topologica che cattura la presenza di confondimento irrisolvibile: due grafi F e F' , con $F' \subseteq F$ e specifiche proprietà di connettività rispetto a T e S , tali che il confondimento non può essere aggiustato con nessuna combinazione di osservabili. La caratterizzazione completa delle *hedge* e il teorema di completezza dell'algoritmo ID nei MAG sono dimostrati in Shpitser e Pearl (2006) — il risultato di completezza per i semi-Markovian causal models è la generalizzazione più potente del do-calculus al caso con variabili latenti.

VII.h Connessione con il problema dell'identificabilità dei LATE in presenza di latenti

La teoria dei MAG ha una connessione diretta con l'interpretazione LATE (Local Average Treatment Effect) degli stimatori a variabile strumentale in presenza di variabili di confondimento latenti — una connessione raramente esplicitata nella letteratura econometrica ma strutturalmente importante.

Quando il DAG del sistema causale contiene un confounder latente U tra il trattamento X e l'esito Y ($X \leftrightarrow Y$ nel MAG proiettato), la stima dell'effetto causale $P(Y \mid \text{do}(X = x))$ richiede uno strumento Z — una variabile che causa X ($Z \rightarrow X$) e non causa direttamente Y (assenza dell'arco $Z \rightarrow Y$ nel DAG). Nel MAG proiettato, lo strumento Z deve soddisfare: Z è adiacente a X (con arco $Z \rightarrow X$), non è adiacente a Y , e non è coinvolto in archi bidirezionali con Y ($Z \leftrightarrow Y$ assente). Questa struttura nel MAG codifica esattamente le tre condizioni di validità dello strumento in forma grafica: la rilevanza ($Z \rightarrow X$), l'esogeneità (assenza di $Z \leftrightarrow Y$, che rappresenterebbe un confondimento latente tra Z e Y), e l'esclusione (assenza dell'arco diretto $Z \rightarrow Y$).

Il LATE di Imbens e Angrist è identificabile in questa struttura come:

$$\text{LATE} = P(Y = 1 \mid \text{do}(X = 1)) - P(Y = 1 \mid \text{do}(X = 0)) \mid (\text{complier subgroup}) = [P(Y = 1 \mid Z = 1) - P(Y = 1 \mid Z = 0)] / [P(X = 1 \mid Z = 1) - P(X = 1 \mid Z = 0)]$$

dove il denominatore è la proporzione di *compliers* — individui per cui Z ha effetto causale su X . La rappresentazione nel MAG chiarisce che il LATE è identificabile perché lo strumento Z "rompe" il confondimento latente $X \leftrightarrow Y$: Z è connesso a Y soltanto attraverso il percorso $Z \rightarrow X \rightarrow Y$, non attraverso archi bidirezionali, e questa struttura garantisce l'identificabilità dell'effetto causale sul sottoinsieme dei compliers.

VIII. La fedeltà e le sue violazioni

L'assunzione di fedeltà è il fondamento della scoperta causale osservazionale. Senza fedeltà, i test di indipendenza condizionale non possono distinguere tra indipendenze strutturali — dovute alla d-separazione nel DAG vero — e indipendenze accidentali — dovute a cancellazioni parametriche che non riflettono la struttura causale.

Violazioni esatte vs. quasi-esatte. Le violazioni di fedeltà "esatte" — coefficienti di percorso che si annullano esattamente — si verificano su un insieme di misura di Lebesgue zero nello spazio dei parametri: una condizione non generica. Tuttavia, le violazioni "quasi-esatte" — coefficienti piccoli ma non zero — non violano formalmente la fedeltà (l'indipendenza statistica non è esatta) ma producono effetti praticamente identici in campioni finiti: il test di indipendenza condizionale non ha la potenza statistica per rilevare un effetto causale molto debole, e il comportamento algoritmico è quello di una violazione di fedeltà effettiva. La dimensione campionaria necessaria per rilevare un effetto causale di ampiezza κ scala come $O(\kappa^{-2})$, rendendo il problema intrattabile per κ piccolo anche con grandi dataset.

Meccanismi tipici. Le violazioni emergono principalmente per: cancellazione di effetti lungo percorsi multipli (un percorso diretto e un percorso indiretto si annullano, come nel sistema $A \rightarrow B, A \rightarrow C \rightarrow B$ con $\beta_{\{AB\}} = -\beta_{\{AC\}} \cdot \beta_{\{CB\}}$); sistemi di feedback omeostatico selezionati evolutivamente (la regolazione spinge i coefficienti verso cancellazioni quasi esatte per mantenere la stabilità); distribuzioni invarianti per simmetrie strutturali specifiche del modello.

Conseguenze algoritmiche. Quando la fedeltà è violata, l'algoritmo PC rimuove erroneamente archi, produce uno scheletro meno connesso del vero, e genera stime del CPDAG sistematicamente distorte verso grafi con meno dipendenze. Non esiste correzione possibile tramite aumento del campione: la violazione quasi-esatta produce lo stesso comportamento dell'esatta nel regime di campione finito realisticamente accessibile. Approcci come il Conservative PC (CPC) di Ramsey et al. (2006) mitigano il problema orientando gli archi con criteri più conservativi, ma non risolvono la questione fondamentale della rilevanza di effetti deboli.

IX. Identificabilità generica: il regime tipico nel caso non gaussiano

Definizione (identificabilità generica). Una struttura causale $Q(M)$ è genericamente identificabile in $\{M_\theta : \theta \in \Theta\}$ se esiste un insieme aperto denso $\Theta_{\text{gen}} \subseteq \Theta$, con complementare di misura di Lebesgue zero, tale che $Q(M_\theta)$ sia identificabile per ogni $\theta \in \Theta_{\text{gen}}$. L'identificabilità generica afferma che la non-identificabilità si verifica soltanto su un insieme sparuto di valori dei parametri — un insieme "trascurabile" — mentre la maggior parte dei modelli è identificabile.

Teorema (Peters e Mooij, 2013). Nel modello lineare bivariato $X \rightarrow Y$ con errori non gaussiani:

$$X := \epsilon_X, Y := \beta \cdot X + \epsilon_Y, \text{ con } \epsilon_X, \epsilon_Y \text{ indipendenti e non gaussiani}$$

la direzione causale è genericamente identificabile: l'unica configurazione in cui la direzione è non-identificabile è quella in cui ϵ_X ed ϵ_Y sono gaussiani, che si verifica su un sottoinsieme di misura zero dello spazio delle distribuzioni degli errori.

La ragione è che la non-gaussianità rompe la simmetria dei modelli gaussiani. Nel regime gaussiano, la distribuzione osservazionale (X, Y) è una gaussiana bivariata completamente

caratterizzata da $(\mu_X, \mu_Y, \sigma^2_X, \sigma^2_Y, \rho)$, e tutte le informazioni causali rilevanti sono contenute in queste cinque quantità di secondo ordine. Ma per ogni distribuzione gaussiana congiunta (X, Y) , esistono coefficienti β, β' tali che sia il modello $X \rightarrow Y$ con parametro β sia il modello $Y \rightarrow X$ con parametro β' riproducono la stessa distribuzione congiunta — la gaussiana è *completa* rispetto alle statistiche del secondo ordine, e questa completezza impedisce l'identificazione della direzione.

Nel regime non gaussiano, le statistiche di ordine superiore (kurtosis, cumulanti di ordine 3 e 4) riflettono la struttura causale in modo asimmetrico tra DAG nella stessa classe di equivalenza di Markov. In particolare, per il modello lineare bivariato con errori non gaussiani, l'indice di LiNGAM:

$$\kappa(\epsilon_X) \cdot \beta^4 \neq \kappa(\epsilon_Y) \text{ per quasi ogni scelta dei parametri}$$

dove κ denota la kurtosis, permette di distinguere la direzione causale da quella inversa tramite i cumulanti di quarto ordine della distribuzione osservata.

Questo risultato è il fondamento teorico di LiNGAM (Linear Non-Gaussian Acyclic Models, Shimizu et al. 2006): la non-gaussianità rompe la simmetria della classe di equivalenza di Markov, rendendo identificabile l'intero DAG anziché il semplice CPDAG. Il costo teorico è l'assunzione di non-gaussianità degli errori: per distribuzioni gaussiane, LiNGAM non fornisce alcuna identificazione aggiuntiva rispetto agli algoritmi standard basati su test di indipendenza condizionale.

X. Sintesi: una geometria dell'ambiguità causale

La teoria delle classi di equivalenza di Markov permette di tracciare con precisione la geometria dello spazio dei modelli causali compatibili con una data distribuzione osservazionale.

Un punto nello spazio dei modelli causali — un DAG con parametri specifici — induce una distribuzione $P(V)$. La fibra della proiezione dalla distribuzione al modello è la classe di equivalenza di Markov: l'insieme di tutti i modelli causali compatibili con quella distribuzione. La fibra non è un punto (a meno di condizioni speciali come la non-gaussianità) ma un insieme esteso, strutturato dal teorema di Verma come i DAG con scheletro e v-strutture identiche.

Questa geometria evidenzia due tipi distinti di ambiguità che si combinano: l'ambiguità di struttura (tra grafi Markov-equivalenti, irriducibile con dati osservazionali) e l'ambiguità di parametri (all'interno di un grafo fissato, in presenza di confondimento latente, irriducibile anche con esperimenti in assenza di assunzioni sulla forma funzionale). I sistemi di intelligenza artificiale addestrati su dati osservazionali non strutturati non identificano nemmeno la classe di equivalenza: apprendono approssimazioni della distribuzione osservazionale — uno strato ancora più esterno rispetto al CPDAG, lontano dal DAG causale per due livelli distinti di ambiguità.

Il quadro della Parte Terza completa così la diagnosi strutturale avviata nella Parte Seconda: non soltanto la distribuzione post-intervento è in generale non determinata dalla

distribuzione osservazionale, ma la struttura grafica qualitativa — il DAG e nemmeno il CPDAG — è in generale non determinata. L'ambiguità causale ha tre strati sovrapposti: ambiguità di parametri (data una struttura, l'effetto causale varia); ambiguità di struttura (data una distribuzione, il CPDAG è il meglio ottenibile); ambiguità di tipo (in presenza di variabili latenti, nemmeno il CPDAG ma solo il PAG è accessibile). Ogni strato ha la propria analisi formale e i propri meccanismi di parziale superamento, come questa Parte ha mostrato sistematicamente.

Parte Quarta: Algoritmi di Scoperta Causale — Possibilità, Limiti e Frontiere Fondamentali

I. Il problema della scoperta causale: inversione dell'inferenza e struttura del programma

La teoria dell'identificabilità sviluppata nelle Parti Seconda e Terza tratta il problema condizionale: dato un grafo G noto, si chiede se le quantità causali di interesse siano calcolabili dalla distribuzione osservazionale. Il problema della scoperta causale è il problema inverso e considerevolmente più ambizioso: non si assume il grafo come noto, e ci si chiede quanto di esso possa essere ricostruito a partire dai soli dati. Questo è il problema che le scienze empiriche affrontano nella pratica ogni volta che non è praticabile progettare esperimenti di intervento — epidemiologia, macroeconomia, astrofisica, genomica, climatologia, psicologia cognitiva, e in modo crescente l'analisi dei sistemi di intelligenza artificiale stessi nella loro dimensione di processi stocastici.

Il programma della scoperta causale è fondato su due assunzioni fondamentali. La prima è la condizione di Markov causale: la distribuzione osservazionale $P(V)$ è Markov rispetto al vero DAG causale G — ogni variabile è condizionalmente indipendente dai suoi non-discendenti dati i suoi genitori nel grafo. La seconda è la condizione di fedeltà: le sole indipendenze condizionali presenti in $P(V)$ sono quelle che la struttura del grafo G implica tramite d-separazione — non vi sono indipendenze accidentali dovute a cancellazioni parametriche esatte.

Sotto queste due assunzioni, la distribuzione $P(V)$ in campioni infiniti determina univocamente il CPDAG, e il problema della scoperta causale si riduce al problema di stimare il CPDAG da dati campionari — cioè da stime della distribuzione osservazionale affette da errore statistico. Ogni algoritmo di scoperta causale è pertanto un algoritmo statistico: riceve un campione D , stima le indipendenze condizionali con test statistici soggetti a errori di tipo I e tipo II, e produce una stima $\hat{G}$ del CPDAG vero. La qualità dell'algoritmo si valuta in termini di consistenza (convergenza al CPDAG corretto in campioni infiniti), velocità di convergenza (dimensione campionaria necessaria con alta probabilità), ed efficienza computazionale.

La Parte è organizzata secondo tre famiglie di algoritmi: quelli basati su test di indipendenza condizionale (PC, PC-stable, FCI), quelli basati su punteggi di verosimiglianza (GES, fGES), e quelli basati su assunzioni funzionali forti (LiNGAM, ANM). Le prime due famiglie

operano sotto le assunzioni di Markov e fedeltà senza vincoli sulla forma funzionale; la terza sfrutta restrizioni sulla distribuzione degli errori o sulla forma delle equazioni per identificare il DAG — non solo la classe di equivalenza — a prezzo di assunzioni più restrittive. Ciascuna famiglia ha limiti computazionali propri e limiti di principio condivisi, sistematicamente separati nell'analisi.

II. L'algoritmo PC: struttura formale e analisi della complessità

L'algoritmo PC — così denominato per i suoi autori principali Peter Spirtes e Clark Glymour, che lo svilupparono nel contesto del progetto TETRAD alla Carnegie Mellon University negli anni Ottanta-Novanta — è il prototipo degli algoritmi di scoperta causale basati su test di indipendenza condizionale. La sua struttura è elegante: si parte dal grafo completo non orientato, si rimuovono gli archi tra coppie condizionalmente indipendenti dato qualche insieme di variabili, si orientano gli archi residui identificando le v-strutture, si propagano le orientazioni con le regole di Meek.

Algoritmo PC — descrizione formale.

Input: variabili $V = \{V_1, \dots, V_n\}$; procedura di test $CI(X, Y \mid Z)$ che restituisce vero (indipendente) o falso (dipendente) sul campione D . Output: CPDAG stimato $\hat{C}$.

Fase 1 — Stima dello scheletro con insiemi di condizionamento crescenti. Si inizializza H come il grafo completo non orientato su V . Per livello di condizionamento $k = 0, 1, 2, \dots$: per ogni coppia adiacente (X, Y) in H , per ogni sottoinsieme $Z \subseteq \text{Adj}(X, H) \setminus \{Y\}$ con $|Z| = k$ — se $CI(X, Y \mid Z) = \text{vero}$, si rimuove l'arco $X - Y$ da H e si registra $\text{Sep}(X, Y) := \text{Sep}(Y, X) := Z$. Si incrementa k finché nessuna coppia adiacente ha abbastanza adiacenti per formare insiemi di dimensione k .

Fase 2 — Orientazione delle v-strutture. Per ogni tripla (X, Z, Y) con $X - Z - Y$ in H e X non adiacente a Y : se $Z \notin \text{Sep}(X, Y)$, si orienta $X \rightarrow Z \leftarrow Y$.

Fase 3 — Propagazione delle orientazioni. Si applicano le regole R1–R4 di Meek fino a convergenza.

Analisi della complessità computazionale. Sia d il grado massimo del vero grafo G . Il numero di test eseguiti scala come:

$$|\text{test}| = O(n^2 \cdot C(d, \lfloor d/2 \rfloor)) \approx O(n^2 \cdot 2^{d/2} / \sqrt{d})$$

Per grafi sparsi con $d = O(1)$, il numero di test è $O(n^2)$ — polinomiale. Per grafi densi con $d = O(n)$, è $O(n^2 \cdot 2^n)$ — esponenziale. La sparsità del grafo è quindi condizione necessaria per la praticabilità computazionale del PC.

Test parametrico per gaussiane: la trasformata di Fisher. Per distribuzioni gaussiane multivariate, il test standard di indipendenza condizionale è basato sulla correlazione parziale $\rho(X, Y \mid Z)$. La trasformata di Fisher del campione $r_{XY|Z}$:

$$z = (1/2) \cdot \log((1 + r_{\{XY|Z\}}) / (1 - r_{\{XY|Z\}}))$$

è approssimativamente distribuita come $N(\varrho\text{-trasformata}, 1/(m - |Z| - 3))$ dove m è la dimensione del campione. Il test $H_0: \varrho(X, Y | Z) = 0$ si esegue confrontando $|z| \cdot \sqrt{(m - |Z| - 3)}$ con la distribuzione $N(0,1)$ al livello di significatività α . La potenza del test decresce con $|Z|$: insiemi di condizionamento grandi richiedono campioni molto più ampi per mantenere la stessa potenza — caratteristica strutturale degli algoritmi basati su test condizionali non eliminabile senza assunzioni parametriche aggiuntive.

Test non parametrici: Kernel-based Conditional Independence. Per distribuzioni non gaussiane, i test di indipendenza condizionale classici non sono applicabili senza errori sistematici. Il KCI (Kernel-based Conditional Independence test) di Zhang et al. (2012) formula il problema come test di indipendenza in uno spazio di Hilbert a kernel riprodotto (RKHS). Dati tre insiemi di variabili X, Y, Z , si definiscono le feature map $\varphi: x \mapsto k_X(\cdot, x)$ e $\psi: y \mapsto k_Y(\cdot, y)$ per kernel k_X e k_Y , e si calcola la statistica:

$$KCI(X, Y | Z) = \mathbb{E}[\varphi(X) \otimes \psi(Y) | Z] - \mathbb{E}[\varphi(X) | Z] \otimes \mathbb{E}[\psi(Y) | Z] \|^2_{\{\text{RKHS}\}}$$

Questa quantità è zero se e solo se $X \perp Y | Z$ sotto condizioni di misurabilità, rendendola un test asintoticamente consistente per indipendenze condizionali arbitrarie. La complessità del KCI è $O(m^3)$ per campioni di dimensione m , proibitiva per grandi dataset; la versione approssimata RCIT di Strobl et al. (2019) usa Random Fourier Features per ridurre la complessità a $O(m \cdot D^2)$ con $D \ll m$.

Consistenza e velocità di convergenza (Kalisch e Bühlmann, 2007). Per distribuzioni gaussiane con grafo sparso G di grado massimo d e forza minima delle dipendenze causali:

$$\kappa = \min\{|\varrho(V_i, V | Z)| : (V_i, V) \text{ adiacenti in } G, Z \subseteq V \setminus \{V_i, V\}, |Z| \leq d\}$$

la dimensione campionaria sufficiente per la consistenza di PC con probabilità $1-\delta$ è:

$$m \geq C \cdot \kappa^{-2} \cdot d^2 \cdot \log(n/\delta)$$

La dipendenza κ^{-2} è la più critica: per sistemi con relazioni causali deboli, il numero di campioni necessari scala come κ^{-2} , crescendo senza bound quando $\kappa \rightarrow 0$. In biologia molecolare, dove le relazioni di regolazione genica hanno tipicamente κ nell'ordine di 0.05–0.20, la formula richiede campioni dell'ordine di 10^4 – 10^6 osservazioni anche per grafi di dimensione moderata.

I. Il problema della scoperta causale: inversione dell'inferenza e struttura del programma

La teoria della identificabilità sviluppata nelle Parti Seconda e Terza tratta il problema condizionale: dato un grafo G noto, si chiede se le quantità causali di interesse siano calcolabili dalla distribuzione osservazionale. Il problema della scoperta causale e il problema

inverso è considerevolmente più ambizioso: non si assume il grafo come noto, e ci si chiede quanto di esso possa essere ricostruito a partire dai soli dati. Questo è il problema che le scienze empiriche affrontano nella pratica ogni volta che non è praticabile progettare esperimenti di intervento e si deve operare su dati osservazionali — epidemiologia, macroeconomia, astrofisica, psicologia cognitiva, genomica, climatologia, e in modo crescente l'analisi dei sistemi di intelligenza artificiale stessi.

I.1 Le due assunzioni fondamentali del programma

Il programma della scoperta causale è fondato su due assunzioni che articolano le condizioni di applicabilità e i limiti dei risultati. La prima è la condizione di Markov causale, già incontrata nella Parte Terza: la distribuzione osservazionale $P(V)$ è Markov rispetto al vero DAG causale G , ossia ogni variabile è condizionalmente indipendente dai suoi non-discendenti dati i suoi genitori nel grafo. La seconda è la condizione di fedeltà: le sole indipendenze condizionali presenti in $P(V)$ sono quelle che la struttura del grafo G implica tramite d-separazione - non vi sono indipendenze accidentali dovute a cancellazioni parametriche esatte.

Condizione di Markov causale (formalizzazione). Un modello causale strutturale M con grafo G soddisfa la condizione di Markov se e solo se, per ogni nodo V_i e per ogni insieme Z che separa V_i dai suoi non-discendenti in G :

$$V_i \perp\!\!\!\perp \text{NonDisc}(V_i) \setminus \text{PA}(V_i) \mid \text{PA}(V_i) \text{ in } P(V)$$

dove $\text{PA}(V_i)$ indica i genitori di V_i in G e $\text{NonDisc}(V_i)$ indica i non-discendenti. Questa condizione è implicata dalla struttura del SCM sotto l'assunzione che le variabili esogene $\{U_i\}$ siano mutualmente indipendenti tra loro (assunzione di esogeneità indipendente). È la condizione più debole possibile che lega la struttura del grafo alla struttura probabilistica: non dice nulla sulla forma delle equazioni strutturali, ma vincola le indipendenze condizionali alla struttura del grafo.

Condizione di fedeltà (formalizzazione). Una distribuzione $P(V)$ è fedele al grafo G se per ogni tripla (A, B, C) :

$$A \perp\!\!\!\perp B \mid C \text{ in } P(V) \implies (A, B \text{ separati da } C \text{ in } G \text{ via d-separazione})$$

La condizione di fedeltà e la condizione di Markov, combinate, creano una corrispondenza biunivoca tra la struttura delle indipendenze condizionali in $P(V)$ e la struttura di d-separazione in G . Sotto queste due assunzioni e in campioni infiniti, la distribuzione $P(V)$ determina univocamente il CPDAG.

Perché la fedeltà è non generica ma necessaria. La fedeltà può essere violata: se due percorsi causali tra A e B si cancellano esattamente (come nel caso di effetti diretti e mediati con segni opposti), allora $A \perp\!\!\!\perp B$ in $P(V)$ pur essendo A e B connessi in G . Tali cancellazioni formano un insieme di misura zero nello spazio dei parametri per qualsiasi modello parametrico continuo — ma non sono impossibili, e in alcuni contesti applicativi (sistemi in retroazione, equilibri di mercato) possono verificarsi quasi esattamente per ragioni sostanziali. I teoremi di consistenza della scoperta causale assumono quindi la fedeltà come precondizione non testabile dai dati, a meno di avere dati interventistici indipendenti.

I.2 Le tre famiglie algoritmiche

Sotto le assunzioni di Markov e fedeltà, esistono tre principali famiglie di algoritmi di scoperta causale, che differiscono nella strategia inferenziale, nelle assunzioni addizionali e nella potenza identificativa:

Famiglia 1: Algoritmi basati su test di indipendenza condizionale (constraint-based).

Questi algoritmi --- di cui PC e FCI sono i rappresentanti principali --- operano testando direttamente le implicazioni di d-separazione del grafo. Per ogni coppia di variabili (X, Y) e ogni possibile insieme di condizionamento Z , si esegue il test $CI(X, Y \mid Z)$ e si usa l'esito per rimuovere o mantenere archi nello scheletro. L'algoritmo produce un CPDAG (o PAG, nel caso FCI con variabili latenti). La solidità teorica è massima: l'algoritmo converge al CPDAG corretto in campioni infiniti sotto sole le assunzioni di Markov e fedeltà, senza ipotesi sulla forma funzionale. Il costo e la complessità computazionale: $O(n^2 \cdot 2^d)$ test nel caso peggiore, dove d è il grado massimo del grafo.

Famiglia 2: Algoritmi basati su punteggi (score-based).

Questi algoritmi - di cui GES (Greedy Equivalence Search) di Chickering (2002) e il prototipo - ottimizzano un punteggio di qualità del modello su tutte le classi di equivalenza di Markov. Il punteggio tipico è il BIC (Bayesian Information Criterion) o il BDeu (Bayesian Dirichlet equivalent uniform). GES esegue una ricerca greedy in due fasi: nella prima (fase di aggiunta) aggiunge archi che migliorano il punteggio, nella seconda (fase di eliminazione) rimuove archi che la loro rimozione migliora il punteggio. La consistenza di GES sotto Markov, fedeltà e gaussianità è dimostrata da Chickering (2002). Il vantaggio rispetto a PC è la maggiore robustezza alle violazioni quasi-esatte della fedeltà: il punteggio BIC è continuo e degrada gradualmente invece di commettere errori di tipo I.

Famiglia 3: Algoritmi basati su assunzioni funzionali (functional-form-based).

Questi algoritmi --- LiNGAM (Shimizu et al., 2006), ANM (Additive Noise Models, Hoyer et al., 2009), PostNonlinear (Zhang e Hyvarinen, 2009) --- sfruttano restrizioni sulla forma delle equazioni strutturali o sulla distribuzione degli errori per identificare il DAG completo, non

solo la classe di equivalenza. LiNGAM assume linearità delle equazioni strutturali e non-gaussianità degli errori: in questo contesto, l'analisi delle componenti indipendenti (ICA) identifica univocamente il DAG. ANM assume rumore additivo indipendente: $X \rightarrow Y$ con $Y = f(X) + N_Y$ e $N_Y \perp\!\!\!\perp X$ è identificabile perché la direzione inversa $X = g(Y) + N_X$ non soddisfa la condizione di indipendenza del rumore per f generica. Questi algoritmi producono risultati più forti (il DAG, non il CPDAG) a prezzo di assunzioni più restrittive che devono essere motivate dal contesto applicativo.

I.3 Il confine fondamentale: osservazione, intervento e controfattuale

Indipendentemente dall'algoritmo utilizzato, la scoperta causale da dati osservazionali incontra un confine fondamentale che i teoremi di impossibilità della Parte Seconda hanno stabilito: il CPDAG, non il DAG, è il massimo identificabile da dati osservazionali in condizioni generali. Per identificare il DAG sono necessari dati interventistici, e anche con questi il numero di interventi necessari ha lower bound logaritmici (teoremi di Eberhardt). I controfattuali individuali rimangono non identificabili in condizioni generali anche con dati interventistici. Questo confine non è un limite degli algoritmi esistenti: è un limite di principio che nessun algoritmo può varcare.

II. L'algoritmo PC: struttura formale, analisi della complessità e prova di consistenza

L'algoritmo PC — denominato per i suoi autori principali Peter Spirtes e Clark Glymour, che lo svilupparono nel contesto del progetto TETRAD alla Carnegie Mellon University tra la fine degli anni Ottanta e i primi anni Novanta — è il prototipo degli algoritmi di scoperta causale basati su test di indipendenza condizionale. La sua struttura concettuale è elegante: si parte dal grafo completo non orientato, si rimuovono gli archi tra coppie condizionalmente indipendenti, si orientano le v-strutture, si propagano le orientazioni con le regole di Meek. La semplicità concettuale non implica semplicità computazionale, come mostrerà l'analisi della complessità.

II.1 Algoritmo PC - descrizione formale

Input: Un insieme $V = \{V_1, \dots, V_n\}$ di variabili, un campione D di m osservazioni, una procedura di test $CI(X, Y \mid Z)$ che restituisce vero o falso, un livello di significatività α .

Output: Un CPDAG stimato G_{hat} .

Fase 1 --- Stima dello scheletro con insiemi di condizionamento crescenti.

Inizializzare H = grafo completo non orientato su V (ogni coppia adiacente).

Inizializzare $Sep(X, Y) = \text{NULL}$ per ogni coppia (X, Y) .

Inizializzare il livello $k = 0$.

Ripetere finché non esiste una coppia adiacente (X, Y) in H con $|Adj(X, H) \setminus \{Y\}| \geq k$:

Per ogni coppia adiacente (X, Y) in H (ordine arbitrario):

Per ogni Z sottoinsieme di $Adj(X, H) \setminus \{Y\}$ con $|Z| = k$:

Se $CI(X, Y \mid Z) = \text{vero}$:

Rimuovere $X - Y$ da H

Registrare $Sep(X, Y) = Sep(Y, X) = Z$

Uscire dal ciclo interno su Z e passare alla prossima coppia

Incrementare $k = k + 1$.

Output della Fase 1: scheletro H e insieme di separatori Sep .

Fase 2 --- Orientazione delle v-strutture.

Per ogni tripla (X, Z, Y) con $X - Z - Y$ in H e X non adiacente a Y in H :

Se Z non appartiene a $Sep(X, Y)$:

Orientare $X \rightarrow Z \leftarrow Y$ (v-struttura: Z è collider non nell'insieme separatore)

Fase 3 --- Propagazione con le regole di Meek. Applicare iterativamente le seguenti regole finché nessuna nuova orientazione emerge:

R1 (Orient away from cycle): Se $A \rightarrow B - C$ e A non adiacente a C , orientare $B \rightarrow C$.

R2 (Acyclicity): Se $A - B$, esiste $A \rightarrow C \rightarrow B$, orientare $A \rightarrow B$.

R3 (Avoid new v-structure): Se $A - C, B - C, A \rightarrow D \leftarrow B$, A non adiacente a B ,

e D adiacente a C , orientare $C \rightarrow D$.

R4 (Avoid new v-structure II): Se $A - B$, $B \rightarrow C \rightarrow D$, $A - D$ ma A non adiacente a C ,

orientare $A \rightarrow B$.

Il CPDAG risultante $G_{\hat{}}$ è il grafo massimale parzialmente orientato compatibile con lo scheletro H , gli insiemi separatori Sep e le regole di orientazione.

II.2 Analisi della complessità computazionale

La complessità dell'algoritmo PC dipende dall'architettura del grafo vero. Sia d il grado massimo del grafo — il massimo numero di vicini di un nodo. Il numero totale di test di indipendenza condizionale eseguiti dall'algoritmo nella Fase 1 è:

$$N_{\text{tests}} \leq \sum_{k=0}^d n^2 * C(d, k) = n^2 * 2^d$$

dove n^2 è il numero di coppie ordinate e $C(d, k)$ è il numero di sottoinsiemi di dimensione k degli adiacenti. Per grafi sparsi con $d = O(1)$, il numero di test è $O(n^2)$ — polinomiale in n . Per grafi densi con $d = O(n)$, il numero di test è $O(n^2 * 2^n)$ — è esponenziale. Il confine critico è $d = O(\log n)$: sopra questa soglia, la complessità diventa super-polinomiale.

Complessità dei singoli test. Per distribuzioni gaussiane, ogni test $CI(X, Y | Z)$ si basa sul calcolo della correlazione parziale:

$$\rho(X, Y | Z) = (\rho(X, Y) - \rho(X, Z) * \rho(Y, Z)) / \sqrt{(1 - \rho(X, Z)^2)(1 - \rho(Y, Z)^2)}$$

dove ρ denota la matrice di correlazione campionaria. Con il calcolo della matrice di correlazione in $O(n^2 * m)$ e ogni test in $O(n^2)$ (per inversione della sottomatrice), la complessità complessiva della Fase 1 è $O(n^4 * m + n^2 * 2^d)$ dove m è la dimensione del campione.

Propagazione degli errori. In campioni finiti, ogni test CI è soggetto ad errori di tipo I (falso positivo: dichiarare indipendenza ove non esiste) e tipo II (falso negativo: non dichiarare indipendenza ove esiste). La propagazione degli errori attraverso le fasi è critica: un arco rimosso erroneamente nella Fase 1 produce uno scheletro sbagliato, che può produrre v-strutture errate nella Fase 2 e orientazioni errate in cascata nella Fase 3. Viceversa, un arco non rimosso erroneamente aumenta la dimensione degli insiemi di condizionamento nei test successivi, riducendo la potenza. Questi effetti di propagazione rendono la stima del CPDAG in campioni finiti un processo in cui gli errori si moltiplicano piuttosto che si compensano.

II.3 Il test di correlazione parziale per variabili gaussiane

Per distribuzioni gaussiane multivariate, il test standard di indipendenza condizionale e il test di correlazione parziale. Dato un campione di dimensione m e un insieme di condizionamento Z di dimensione $|Z| = q$, la correlazione parziale campionaria $\rho_{\text{hat}}(X, Y | Z)$ e la correlazione di X e Y nei residui della regressione lineare di X su Z e di Y su Z . Sotto l'ipotesi nulla $H_0: \rho(X, Y | Z) = 0$ (vera indipendenza condizionale) e sotto normalità, la statistica:

$$t = \sqrt{m - q - 3} * (1/2) * \log((1 + \rho_{\text{hat}}) / (1 - \rho_{\text{hat}}))$$

(trasformata z di Fisher) segue asintoticamente una distribuzione normale standard $N(0,1)$, e esattamente una distribuzione t di Student con $m - q - 3$ gradi di libertà. Il test rifiuta l'ipotesi nulla al livello α se $|t| > t_{\{\alpha/2, m-q-3\}}$.

Osservazione critica. La potenza del test dipende dalla dimensione campionaria m , dalla forza della dipendenza residua (la correlazione parziale vera ρ), e dalla dimensione q dell'insieme di condizionamento. Per ρ piccolo (dipendenza causale debole) e q grande (insiemi di condizionamento grandi, necessari per grafi densi), la potenza decade rapidamente. Formalmente, per rilevare una dipendenza con correlazione parziale $\kappa > 0$ a livello di potenza $1 - \beta$:

$$m \geq (z_{\{\alpha/2\}} + z_{\beta})^2 / \kappa^2 + q + 3$$

Questa stima - precursore del teorema di Kalisch-Buhlmann che seguirà - mostra che la dimensione campionaria scala come κ^{-2} : per $\kappa = 0.05$ (dipendenze deboli, tipiche in biologia molecolare), servono almeno $m \sim 400 * q$ campioni per mantenere la potenza. Poiché q può essere fino a d (il grado massimo), il costo campionario totale scala come $\kappa^{-2} * d$.

II.4 Prova di consistenza dell'algoritmo PC in campioni infiniti

Teorema (Spirtes, Glymour, Scheines, 1993 --- consistenza di PC). Sia P una distribuzione osservazionale su un insieme V di n variabili, Markov e fedele rispetto a un DAG causale G . Sia il test CI oracolo, cioè $CI(X, Y | Z) = \text{vero}$ se e solo se $X \perp\!\!\!\perp Y | Z$ in P . Allora l'algoritmo PC produce il CPDAG $C(G)$ --- la rappresentazione della classe di equivalenza di Markov di G .

Prova. La prova procede in tre lemmi corrispondenti alle tre fasi dell'algoritmo.

Lemma 1 (Fase 1: correttezza dello scheletro). Sotto le assunzioni di Markov e fedeltà con test CI oracolo, l'algoritmo PC produce lo scheletro corretto di G , ovvero: $X - Y$ in H se e solo se X e Y sono adiacenti in G .

Prova del Lemma 1. ($\Rightarrow$) Se X e Y sono adiacenti in G , allora per la condizione di Markov non possono essere d-separati da nessun insieme Z che non contenga archi diretti tra X e Y . In particolare, non esiste Z tale che $X _ | _ Y | Z$ in P (per la fedeltà: d-separazione è necessaria per l'indipendenza). Dunque il test CI non troverà mai indipendenza per la coppia (X, Y) , e l'arco non verrà rimosso.

($\Leftarrow$) Se X e Y non sono adiacenti in G , allora per il teorema di Verma (1987) esiste un insieme Z che d-separa X e Y in G , e tale Z è contenuto nella frontiera di Markov di X o di Y (insieme dei genitori, figli e co-genitori in G). Per la condizione di Markov, $X _ | _ Y | Z$ in P . L'insieme Z ha cardinalità al più $\max(|PA(X)|, |PA(Y)|) \leq d$ (il grado massimo del grafo). Quindi, quando l'algoritmo PC raggiunge il livello $k = |Z|$, eseguirà il test $CI(X, Y | Z)$, otterrà vero, e rimuoverà l'arco. QED Lemma 1.

Lemma 2 (Fase 2: correttezza delle v-strutture). Sotto le assunzioni del Lemma 1 e con scheletro corretto, la Fase 2 orienta esattamente le v-strutture di G , ovvero: $X \rightarrow Z \leftarrow Y$ è una v-struttura in G se e solo se $X - Z - Y$ in H , X non adiacente a Y , e Z non appartiene a $Sep(X, Y)$.

Prova del Lemma 2. ($\Rightarrow$) Se $X \rightarrow Z \leftarrow Y$ è una v-struttura in G , allora X e Y non sono adiacenti (per definizione di v-struttura). Per la condizione di Markov, Z non d-separa X da Y per nessun Z in G (perché Z è un collider nel percorso $X \rightarrow Z \leftarrow Y$, e aprire un collider richiede condizionare su di esso o suoi discendenti). Dunque il test $CI(X, Y | Z)$ doveva risultare falso ogni volta che Z è incluso nell'insieme di condizionamento, e $Sep(X, Y)$ deve essere un insieme che non include Z .

($\Leftarrow$) Supponiamo $X - Z - Y$ in H , X non adiacente a Y , e Z non in $Sep(X, Y)$. Per la correttezza dello scheletro (Lemma 1), Z è adiacente a X e Y in G . Per le condizioni di Markov e fedeltà, $Sep(X, Y)$ deve essere un insieme che d-separa X e Y . Poiché Z è adiacente a X e Y ma non in $Sep(X, Y)$, la tripla (X, Z, Y) deve essere una v-struttura in G (se non lo fosse, Z sarebbe un non-collider nel percorso $X - Z - Y$, e includere Z in Sep avrebbe bloccato il percorso, contraddicendo fedeltà). QED Lemma 2.

Lemma 3 (Fase 3: completezza delle orientazioni di Meek). Le regole R1-R4 di Meek producono esattamente le orientazioni degli archi che sono invarianti in tutti i DAG della classe di equivalenza di Markov di G .

Prova del Lemma 3. Meek (1995) dimostra che le regole R1-R4 sono complete per la proprietà di completezza dell'orientazione parziale: un arco $X - Y$ nel CPDAG parzialmente orientato deve essere orientato come $X \rightarrow Y$ se e solo se in ogni DAG della classe di equivalenza l'arco va da X a Y . Le regole sono:

R1 garantisce aciclicità: se $A \rightarrow B \leftarrow C$ e A non adiacente a C , $B \rightarrow C$ deve essere orientato per evitare che il ciclo $A \rightarrow B \leftarrow C$ implichi l'esistenza di un ciclo con archi $B \rightarrow A$ non orientabile.

R2 evita cicli diretti: se $A \rightarrow C \rightarrow B$ ma $A \leftarrow B$, $A \rightarrow B$ deve essere orientato per evitare il ciclo $B \rightarrow A \rightarrow C \rightarrow B$.

R3 e R4 evitano nuove v-strutture implicite che non sarebbero giustificate dagli insiemi separatori.

La completezza di R1-R4 - il fatto che non siano necessarie regole addizionali - segue dal teorema di caratterizzazione dei CPDAG di Chickering (1995): ogni CPDAG si ottiene orientando esattamente gli archi protetti di un grafo non orientato compatibile, e R1-R4 identificano esattamente questi archi. QED Lemma 3 e Teorema principale.

III. Il teorema di Kalisch-Buhlmann: consistenza in alta dimensione

Il teorema di consistenza in campioni infiniti di Spirtes, Glymour e Scheines garantisce solo la convergenza asintotica. Per le applicazioni in campioni finiti - in particolare nel regime di alta dimensione $n \gg 1$ con m campioni - è essenziale disporre di stime esplicite della dimensione campionaria necessaria per la consistenza con alta probabilità. Questo è il contributo centrale di Kalisch e Buhlmann (2007), che hanno esteso la consistenza dell'algoritmo PC al regime $n \gg m$ (più variabili che osservazioni) sotto l'assunzione di sparsità del grafo.

III.1 Enunciato del teorema di Kalisch-Buhlmann

Ipotesi. Siano verificate le seguenti condizioni:

(KB1) P è una distribuzione gaussiana multivariata su n variabili.

(KB2) G è il grafo causale vero, con grado massimo d (numero massimo di vicini per nodo).

(KB3) P soddisfa la condizione di Markov e fedeltà rispetto a G .

(KB4) $\kappa > 0$ e la forza minima delle dipendenze causali:

$$\kappa = \min \{ |\rho(V_i, V_j | Z)| : (V_i, V_j) \text{ adiacenti in } G, Z \text{ sottoinsieme di } V \setminus \{V_i, V_j\}, |Z| \leq d \}$$

(KB5) δ in $(0, 1)$ e il livello di probabilità desiderato.

Teorema (Kalisch e Buhlmann, 2007). Sotto le ipotesi (KB1)-(KB5), l'algoritmo PC con test di correlazione parziale e livello di significatività α_m che soddisfa $\alpha_m \rightarrow 0$ e $m * \alpha_m \rightarrow \infty$ stima il CPDAG corretto con probabilità almeno $1 - \delta$ se la dimensione campionaria soddisfa:

$$m \geq C * d^2 * \log(n / \delta) / \kappa^2$$

dove C è una costante assoluta che dipende solo dalla distribuzione del test statistico (per il test di Fisher-z gaussiano, $C \sim 8/(1 - \tanh^2(\kappa))^2 \sim 8$ per κ piccolo).

III.2 Analisi strutturale del bound

La struttura della stima $m \geq C * d^2 * \log(n / \delta) / \kappa^2$ merita un'analisi attenta dei tre fattori:

Dipendenza logaritmica in n . Il fattore $\log(n / \delta)$ è ottimale per il problema di stima del grafo. Formalmente, qualsiasi algoritmo consistente per la scoperta causale deve eseguire almeno $\Omega(n^2)$ test, e ogni test richiede almeno $\Omega(\log(n / \delta) / \kappa^2)$ campioni per mantenere la probabilità di errore sotto δ / n^2 (per una union bound sui test). Il fattore $\log(n)$ nella stima di Kalisch-Buhlmann è quindi il minimo possibile. Questo significa che nel regime high-dimensional con n crescente ma $m = \Omega(d^2 * \log n)$, l'algoritmo PC rimane consistente anche quando $n \gg m$, a condizione che il grafo sia sparso e κ sia limitato inferiormente.

Dipendenza quadratica in d . Il fattore d^2 riflette il costo dei test di indipendenza condizionale su insiemi di dimensione d . Per ogni coppia adiacente (V_i, V_j) , il test più "difficile" è quello che condiziona sull'intero vicinato di dimensione d . La correlazione parziale $\rho(V_i, V_j | Z)$ con $|Z| = d$ si stima con una regressione su d variabili, che richiede $m \geq d + 3$ campioni per avere il test statisticamente valido ($d + 3$ gradi di libertà). La dipendenza d^2 invece di d lineare emerge perché la stima della correlazione parziale su insieme di dimensione d ha varianza proporzionale a $1/(m - d - 3)^2 \sim 1/m^2$ per $m \gg d$, e la stima del bound richiede quindi $m \geq C * d^2 * \log(n/\delta) / \kappa^2$.

Dipendenza κ^{-2} e il limite critico. Lai con relazioni causali deboli (κ piccolo), il numero di campioni scala come κ^{-2} , crescendo senza bound quando $\kappa \rightarrow 0$. In biologia molecolare, dove le relazioni di regolazione genica sono tipicamente deboli (κ dell'ordine di 0.05 — 0.20), la stima richiede campioni dell'ordine di 10^4 — 10^6 osservazioni anche per grafi di dimensione moderata. Per $\kappa = 0.05$ e $d = 3$, $n = 1000$ nodi, $\delta = 0.05$:

$$m \geq 8 * 9 * \log(20000) / 0.0025 \sim 8 * 9 * 9.9 / 0.0025 \sim 285.000 \text{ campioni}$$

Un numero di campioni dell'ordine di 285.000 è superiore di due ordini di grandezza a quanto tipicamente disponibile negli studi genomici (100 — 1000 campioni per migliaia di geni). Questo illustra in modo quantitativo il problema pratico della scoperta causale in alta dimensione.

III.3 Schema della prova del teorema di Kalisch-Buhlmann

La prova si svolge in due passi: prima si dimostra che il test di correlazione parziale identifica correttamente le dipendenze e indipendenze condizionali con probabilità alta, poi si usa questa garanzia per applicare il teorema di consistenza di PC con test oracolo.

Passo 1: Controllo uniforme degli errori di test. Per ogni coppia (V_i, V_j) e ogni insieme Z con $|Z| \leq d$, il test di Fisher-z rifiuta correttamente l'ipotesi nulla $\rho(V_i, V_j | Z) = 0$ con probabilità almeno $1 - \alpha_m$ se la correlazione parziale vera è zero, e non la rifiuta con probabilità almeno $1 - \beta_m$ se la correlazione parziale vera ha valore assoluto almeno κ . Le probabilità di errore α_m e β_m possono essere rese simultaneamente piccole per m sufficientemente grande.

Il numero totale di test è al più $n^2 * C(n, d) \sim n^{d+2} / d!$ nella Fase 1. Per controllare uniformemente tutti i test con probabilità complessiva di errore almeno $1 - \delta$, è necessario che la probabilità di errore per ogni singolo test sia al più $\delta / (n^{d+2})$. Per il test di Fisher-z, questo richiede:

$$m \geq C * \log(n^{d+2} / \delta) / \kappa^2 = C * (d+2) * \log(n) / \kappa^2 + C * \log(1/\delta) / \kappa^2$$

che è $O(d * \log(n) / \delta) / \kappa^2$. Il termine d^2 nella stima finale emerge perché i test con $|Z| = d$ richiedono $d + 3$ gradi di libertà, e la dimensione campionaria effettiva è $m - d - 3$; richiedere $m - d - 3 \geq C * \log(n/\delta) / \kappa^2$ con d potenzialmente $O(\sqrt{m})$ introduce un termine quadratico nel caso peggiore.

Passo 2: Applicazione del teorema di consistenza. Se tutti i test producono risultati corretti --- con probabilità almeno $1 - \delta$ per m sufficientemente grande come nel Passo 1 --- allora la condizione del teorema di Spirtes-Glymour-Scheines (test CI oracolo) è soddisfatta con alta probabilità, e il CPDAG stimato coincide con il CPDAG corretto. Combinando Passo 1 e Passo 2 si ottiene il teorema di Kalisch-Buhlmann.

III.4 Il regime high-dimensional: $n \gg m$ con sparsità

Il contributo più importante del teorema di Kalisch-Buhlmann è che permette $n \gg m$: il numero di variabili può essere esponenzialmente più grande della dimensione campionaria, a condizione che il grafo sia sparso ($d = O(\log m)$ o più piccolo). Questo è il regime tipico della bioinformatica ($n = 10.000 \text{ --- } 30.000$ geni, $m = 100 \text{ --- } 1000$ campioni), dell'analisi delle reti neurali, e della scoperta causale in sistemi ad alta dimensionalità.

Formalmente, la condizione $m \geq C * d^2 * \log(n / \delta) / \kappa^2$ permette n esponenziale in m : per m campioni con d e κ fissati, la dimensione del grafo può essere fino a $n \sim \exp(m * \kappa^2 / (C * d^2))$ e la consistenza è garantita. Con $m = 1000$, $d = 3$, $\kappa = 0.20$ (relazioni moderate), $C = 8$:

$$n_{\max} \sim \exp(1000 * 0.04 / (8 * 9)) = \exp(40/72) \sim \exp(0.56) \sim 1.74$$

Questo risultato apparentemente deludente ($n_{\max} \sim 2$ per $m = 1000$) illustra come $\kappa = 0.20$ sia già una relazione relativamente debole che richiede moltissimi campioni. Per $\kappa = 0.50$ (relazioni moderate-forti):

$$n_{\max} \sim \exp(1000 * 0.25 / (8 * 9)) = \exp(250/72) \sim \exp(3.47) \sim 32$$

Anche in questo caso $n_{\max} = 32$ con $m = 1000$ campioni, molto lontano dai regimi high-dimensional della bioinformatica. La conseguenza pratica è che i risultati teorici di consistenza di Kalisch-Buhlmann si applicano con relazioni causali di forza moderata o forte ($\kappa \geq 0.3$ circa) per avere regimi di applicabilità realistica.

III.5 Estensioni e varianti: PC-stable e alta dimensione

Il problema dell'ordine e PC-stable. L'algoritmo PC classico è sensibile all'ordine in cui le coppie sono processate nella Fase 1: se la rimozione dell'arco (X, Y) cambia l'insieme degli adiacenti di X , i test successivi su coppie che coinvolgono X usano insiemi di condizionamento diversi da quelli che sarebbero stati usati se le coppie fossero state processate in ordine diverso. In campioni finiti — dove diversi test su stesso scheletro producono risultati marginalmente diversi — questo crea instabilità: due correnti del campione con ordine di processing diverso possono produrre CPDAG diversi per lo stesso grafo vero.

Definizione (PC-stable, Colombo e Maathuis, 2014). La versione PC-stable elimina la dipendenza dall'ordine modificando la Fase 1: in ogni ciclo del livello k , si usano gli insiemi di adiacenti del grafo all'inizio del ciclo (cioè prima di rimuovere qualsiasi arco nel ciclo corrente), e le rimozioni avvengono solo alla fine del ciclo. Questo garantisce che, per un

dato livello k , tutti i test usano lo stesso insieme di adiacenti — rendendo il risultato invariante all'ordine di processing.

Formalmente, sia H_k il grafo all'inizio del livello k . PC-stable esegue tutti i test $CI(X, Y \mid Z)$ con Z sottoinsieme di $Adj(X, H_k)$ (non di $Adj(X, H_{corrente})$) e rimuove gli archi solo dopo aver completato tutti i test del livello k . Il teorema di consistenza si estende direttamente a PC-stable: le garanzie di correttezza e complessità campionaria restano identiche, con il vantaggio della invarianza all'ordine.

Algoritmi locali: MMPC e Markov blanket. Per grafi ad alta dimensione con $n \gg 1000$, la stima del CPDAG globale può essere computazionalmente proibitiva anche per grafi sparsi. Gli algoritmi che stimano il CPDAG locale intorno a una variabile di interesse --- il cosiddetto Markov blanket di V_i --- sono più efficienti. Il Markov blanket $MB(V_i)$ è il minimo insieme di variabili che rende V_i condizionalmente indipendente dal resto:

$$MB(V_i) = PA(V_i) \cup CH(V_i) \cup PA(CH(V_i))$$

dove PA e CH denotano genitori e figli nel DAG vero. L'algoritmo MMPC (Max-Min Parents and Children) di Tsamardinos et al. (2003) stima il Markov blanket in tempo $O(n * d * m)$ dove m è la dimensione del campione, significativamente più efficiente del PC globale per grafi ad alta dimensione. Le garanzie di consistenza di MMPC seguono dalla stessa analisi di Kalisch-Buhlmann applicata localmente.

III. PC-stable, FCI e il regime ad alta dimensione

PC-stable (Colombo e Maathuis, 2014). L'algoritmo PC classico è sensibile all'ordine in cui le coppie sono processate nella Fase 1: la rimozione di un arco modifica gli insiemi di adiacenti, influenzando i test successivi in modo dipendente dall'ordine. PC-stable elimina questa dipendenza: in ogni ciclo di livello k , si usano gli insiemi di adiacenti del grafo prima di rimuovere qualsiasi arco nel ciclo corrente, e tutte le rimozioni avvengono solo alla fine del ciclo. Questo rende l'algoritmo order-independent e migliora la riproducibilità tra implementazioni diverse senza alterare la complessità asintotica.

L'algoritmo FCI: scoperta causale con variabili latenti. L'algoritmo FCI (Fast Causal Inference) di Spirtes, Meek e Richardson è la generalizzazione di PC che rilascia l'ipotesi di sufficienza causale e gestisce la presenza di variabili latenti, producendo una stima del PAG piuttosto che del CPDAG. La struttura dell'algoritmo ha la stessa ossatura del PC — stima dello scheletro tramite test di indipendenza condizionale, orientazione delle v-strutture, propagazione — ma con importanti differenze strutturali.

Nella Fase 1 del FCI, gli insiemi di condizionamento possono includere variabili non adiacenti nel grafo corrente (i possible-ancestors), rendendola più costosa del PC. Nella Fase 2, si inizializzano tutti gli estremi degli archi come cerchi ($\circ$) e si orientano le v-strutture non

ambigue. Nella Fase 3, si applicano le regole di orientazione R1–R10 di Zhang (2008), estensioni delle regole di Meek per i grafi ancestrali parziali, che gestiscono i tre marcatori agli estremi (freccia, cerchio, lineetta) e le incertezze molteplici della struttura PAG.

Teorema (Spirtes, Meek, Richardson — consistenza di FCI). Sotto le assunzioni di Markov e fedeltà per il MAG (il grafo marginale su V ottenuto dal DAG completo che include le variabili latenti), l'algoritmo FCI converge in campioni infiniti al PAG corretto del MAG di G rispetto alle variabili osservate.

In campioni finiti, l'instabilità di FCI è marcatamente superiore a quella di PC: l'incertezza sui marcatori degli archi — tre possibilità invece di due — si traduce in varianza maggiore delle stime, e gli errori nella Fase 1 si propagano in modo più complesso. La versione RFCI (Really Fast Causal Inference) di Colombo, Maathuis, Kalisch e Richardson (2012) riduce significativamente il numero di test limitando gli insiemi di condizionamento, accettando una riduzione della completezza formale in cambio di praticabilità computazionale su grafi di grandi dimensioni.

PC ad alta dimensione e algoritmi locali. Nel regime $n \gg m$, la versione high-dimensional del PC restringe gli insiemi di condizionamento a dimensione al più $k_{\max} = \lfloor \log(m) \rfloor$, riducendo la complessità e permettendo l'applicazione in regimi dove n è dell'ordine di 10^3 o 10^4 con m dell'ordine di 10^2 – 10^3 , rilevante per la genomica e il neuroimaging. L'algoritmo MMPC (Max-Min Parents and Children) di Tsamardinos et al. stima il Markov blanket $MB(Y)$ — il sottoinsieme minimo tale che $Y \perp V \setminus MB(Y) \cup \{Y\} \mid MB(Y)$ — con complessità $O(n \cdot |MB(Y)|)$ test, molto inferiore al costo globale $O(n^2 \cdot 2^d)$ quando $|MB(Y)| \ll n$.

IV. L'algoritmo GES: scoperta causale basata su punteggi di verosimiglianza

L'algoritmo GES (Greedy Equivalence Search) di Chickering (2002) è il prototipo degli algoritmi score-based per la scoperta del CPDAG. Opera in due fasi di ricerca greedy nello spazio dei CPDAG, anziché testare singole indipendenze locali e costruire il grafo in modo constraint-based.

Fase 1 (Forward) — Inserzione di archi: partendo dal grafo vuoto, a ogni passo si inserisce la modifica del CPDAG che massimizza il guadagno nel punteggio. Si continua fino a convergenza. Fase 2 (Backward) — Eliminazione di archi: partendo dal CPDAG prodotto dalla Fase 1, a ogni passo si elimina l'arco la cui rimozione massimizza il guadagno. Si continua fino a convergenza.

Il punteggio tipicamente usato è il BIC (Bayesian Information Criterion):

$$\text{BIC}(G, D) = \log P(D \mid \hat{G}_{\text{MLE}}) - (|G|/2) \cdot \log m$$

dove $|G|$ è il numero di parametri liberi e m la dimensione del campione. Per modelli gaussiani lineari, il BIC è calcolabile in forma chiusa tramite le correlazioni campionarie, e la proprietà di decomposabilità — $\text{BIC}(G) = \sum_i \text{BIC}_{\text{locale}}(V_i \mid \text{PA}(V_i))$ — rende ogni aggiornamento del punteggio computazionalmente efficiente.

Consistenza di GES (Chickering, 2002). Sotto le assunzioni di Markov, fedeltà e decomposabilità del punteggio con la proprietà di score equivalence (DAG Markov-equivalenti ricevono lo stesso punteggio), GES converge al CPDAG corretto con probabilità che tende a 1 al crescere del campione. La consistenza di GES è più robusta di quella di PC rispetto a violazioni quasi-esatte della fedeltà: GES ottimizza un criterio globale piuttosto che testare indipendenze locali, e piccoli effetti causali che causerebbero falsi positivi nel PC sono incorporati come contributi positivi al punteggio.

fGES (Fast GES, Ramsey et al., 2017). La versione fGES usa strutture dati efficienti con lazy evaluation per ridurre la complessità della fase Forward da $O(n^3)$ a $O(n^2 \cdot |E|)$ nel numero di archi del grafo risultante, rendendola praticabile per grafi con $n = 10^3$ – 10^4 variabili e applicabile su dataset genomici, in neuroimaging funzionale, e in applicazioni climatologiche.

V. LiNGAM: identificabilità tramite non-gaussianità degli errori

L'algoritmo LiNGAM (Linear Non-Gaussian Acyclic Model) di Shimizu, Hoyer, Hyvärinen e Kerminen (2006) produce un risultato qualitativamente distinto da PC e GES: identifica il DAG orientato completo, non solo la classe di equivalenza, a prezzo di assunzioni sulla distribuzione degli errori.

Il modello LiNGAM. Le equazioni strutturali sono lineari: $V = B \cdot V + E$, dove B è una matrice $n \times n$ triangolare inferiore nell'ordinamento topologico del DAG con $B_{ii} \neq 0$ se e solo se $V \in \text{PA}(V_i)$, e $E = (E_1, \dots, E_n)$ è il vettore degli errori con componenti mutuamente indipendenti e distribuzioni non gaussiane (almeno $n-1$ componenti non gaussiane). La soluzione del sistema è $V = (I - B)^{-1}E =: A \cdot E$, dove $A = (I - B)^{-1}$ — il vettore V è una miscelazione lineare tramite A di sorgenti indipendenti E , esattamente il modello dell'Analisi delle Componenti Indipendenti (ICA).

Teorema (Shimizu et al., 2006 — identificabilità di LiNGAM). Sotto le assunzioni di aciclicità, linearità delle equazioni strutturali, indipendenza mutua degli errori, e non-gaussianità di almeno $n-1$ componenti di errore, il DAG causale è identificabile — non solo la classe di equivalenza di Markov, ma il DAG orientato con tutte le direzioni degli archi.

La dimostrazione si basa sul teorema fondamentale dell'ICA (Comon, 1994): se $V = A \cdot E$ con E sorgenti indipendenti non tutte gaussiane, allora A è identificabile a meno di permutazioni delle colonne e riscalamento. Poiché $A = (I - B)^{-1}$ ha struttura triangolare nell'ordinamento topologico, la permutazione delle colonne corrisponde a una permutazione dell'ordinamento topologico — identificabile dall'unicità dell'ordinamento compatibile con la struttura triangolare di B .

Perché la gaussianità non consente l'identificazione. Se tutti gli errori sono gaussiani indipendenti, allora V è un vettore gaussiano multivariato. La distribuzione gaussiana multivariata è invariante rispetto a trasformazioni ortogonali: per qualsiasi matrice ortogonale O , il modello $V = A \cdot O^{-1} \cdot (O \cdot E)$ è osservazionalmente equivalente al modello $V = A \cdot E$, con A sostituito da $A \cdot O^{-1}$ e E sostituito da $O \cdot E$. Questa ambiguità rotazionale rende impossibile

identificare la matrice di miscelazione A e quindi il DAG causale dal solo modello gaussiano. La non-gaussianità rompe questa simmetria: le statistiche di ordine superiore (kurtosis, cumulanti di ordine 3 e 4) sono sensibili alla direzione degli assi e permettono di identificare A in modo univoco.

Algoritmo LiNGAM — implementazione pratica. L'implementazione standard procede in tre passi. Nel Passo 1, si applica un algoritmo ICA ai dati V per stimare $\hat{A}$ e quindi $\hat{W} = \hat{A}^{-1}$. Nel Passo 2, si trova la permutazione π delle righe di $\hat{W}$ che rende la matrice $P_{\pi} \cdot \hat{W}$ più vicina a una triangolare inferiore con diagonale nulla, risolvendo: $\hat{\pi} = \operatorname{argmin}_{\pi} \sum_i (P_{\pi} \cdot \hat{W})_{ii}^2$, trattabile tramite l'algoritmo ungherese in tempo $O(n^3)$. Nel Passo 3, si stima $B = I - (P_{\pi} \cdot \hat{W})^{-1}$ e si determina il DAG dalla struttura di non-zeri di B .

DirectLiNGAM (Shimizu et al., 2011). Una versione alternativa evita la stima ICA — instabile per grafi con n grande — ed esegue la scoperta causale per passi iterativi: a ogni passo si identifica la variabile "più esogena" — quella la cui regressione lineare sulle altre produce residui massimamente indipendenti dalle covariate — e la si rimuove dal sistema. Questa versione è più stabile numericamente e scala meglio ad alta dimensione. Tra i limiti strutturali di LiNGAM: la linearità è la restrizione più severa; la non-gaussianità può essere violata per somma di molti contributi indipendenti; la stima ICA richiede campioni di dimensione $O(n^2)$ – $O(n^3)$ per convergenza affidabile.

VI. Additive Noise Models: identificabilità nel caso bivariato e multivariato

I modelli a rumore additivo (ANM) generalizzano LiNGAM al caso non lineare:

$$V_i := f_i(\text{PA}(V_i)) + E_i$$

dove f_i è una funzione arbitraria dei genitori causali e E_i è un termine di errore additivo indipendente da $\text{PA}(V_i)$. L'indipendenza del termine di errore dai predittori è la proprietà strutturale che permette l'identificazione della direzione causale.

Teorema di identificabilità degli ANM nel caso bivariato (Hoyer et al., 2009; Peters et al., 2012). Sia $X \rightarrow Y$ il vero modello causale bivariato: $Y = f(X) + E_Y$ con $E_Y \perp X$. La condizione di indipendenza nella direzione inversa — $X = g(Y) + E_X$ con $E_X \perp Y$ — non può essere soddisfatta per nessuna funzione g e nessuna distribuzione di E_X , eccetto per famiglie di misura zero nello spazio delle funzioni, quando f non è lineare o E_Y non è gaussiano.

La condizione di non-identificabilità inversa è soddisfatta esattamente quando f , E_Y soddisfano l'equazione differenziale di Darmon-Skitović:

$$\partial^2 / \partial x \partial y \log p_{\{X,Y\}}(x, y) = 0$$

che caratterizza esattamente le distribuzioni gaussiane con funzioni lineari. Per ogni altra coppia (f, E_Y) generica, la struttura ANM è asimmetrica: la direzione causale corretta $X \rightarrow Y$ produce residui $E_Y = Y - f(X)$ indipendenti da X , mentre la direzione inversa non può produrre residui indipendenti da Y per nessuna regressione.

Algoritmo ANM pratico per il caso bivariato. Nel Passo 1, si stima f tramite regressione non parametrica di Y su X (Gaussian process regression, random forest, kernel ridge regression). Nel Passo 2, si calcolano i residui $R_X = Y - f(X)$. Nel Passo 3, si testa $H_0: R_X \perp X$ tramite un test di indipendenza (HSIC). Nel Passo 4, si ripete nella direzione inversa stimando $\hat{g}$ tramite regressione di X su Y e testando $R_Y \perp Y$. Nel Passo 5, la direzione causale è quella per cui il test non rifiuta l'indipendenza.

La statistica HSIC (Gretton et al., 2007). Per due variabili X e Y con kernel k_X e k_Y :

$$\text{HSIC}(X, Y) = \mathbb{E}[k_X(X, X') \cdot k_Y(Y, Y')] - 2\mathbb{E}[k_X(X, X') \cdot k_Y(Y, Y'')] + \mathbb{E}[k_X(X, X'')] \cdot \mathbb{E}[k_Y(Y, Y'')]$$

dove le aspettative sono calcolate su copie indipendenti di (X, Y) . HSIC è zero se e solo se $X \perp Y$ sotto condizioni di misurabilità del kernel, ed è sempre non-negativo. L'HSIC campionario è calcolabile in $O(m^2)$ tramite le matrici kernel centralizzate, e la distribuzione nulla è approssimata tramite permutazioni.

Estensione multivariata degli ANM. Per sistemi con più variabili, Peters et al. (2014) dimostrano che il DAG causale completo è identificabile dagli ANM — non solo la classe di equivalenza — sotto le assunzioni di non-linearità generica delle funzioni f_i e non-gaussianità generica degli errori. L'algoritmo procede identificando iterativamente la variabile "più causata" — quella con residui massimamente indipendenti dai predittori — e costruisce l'ordinamento topologico in ordine inverso.

VII. Test di indipendenza condizionale kernel-based: trattazione sistematica

I test kernel-based meritano una trattazione sistematica per la crescente importanza nella scoperta causale su dati non gaussiani. La struttura generale di un test kernel-based di indipendenza condizionale $X \perp Y \mid Z$ funziona in tre passi: si neutralizza la dipendenza da Z tramite proiezione negli RKHS condizionali; si misura la dipendenza residua X - Y ; si testa se questa dipendenza residua è statisticamente significativa.

La neutralizzazione avviene tramite la regressione negli RKHS: si proiettano le feature map di X e Y sullo spazio ortogonale alle feature di Z . L'indipendenza condizionale $X \perp Y \mid Z$ è equivalente, nel senso RKHS, alla condizione che i residui RKHS di X e Y siano non correlati nello spazio di Hilbert.

La statistica KCI e la distribuzione nulla. La statistica del test KCI è:

$$\text{KCI}_n = (1/n^2) \cdot \text{Tr}(K_X^c \cdot K_Y^c)$$

dove K_X^c e K_Y^c sono le matrici kernel centralizzate condizionalmente. Sotto $H_0: X \perp Y \mid Z$, la statistica è approssimativamente distribuita come una miscela di chi-quadrato pesata dagli autovalori del kernel condizionale centralizzato, approssimata tramite il metodo del matching dei momenti.

RCIT e Random Fourier Features. Per dataset grandi ($m > 10^4$), la complessità $O(m^3)$ del KCI è proibitiva. Il RCIT di Strobl et al. (2019) approssima le feature map RKHS tramite Random Fourier Features:

$$\varphi_D(x) = (1/\sqrt{D}) \cdot [\cos(\omega_1^T x + b_1), \dots, \cos(\omega_D^T x + b_D)]$$

dove $\omega_j \sim p(\omega)$ è la distribuzione spettrale del kernel e $b_j \sim \text{Uniform}(0, 2\pi)$. Con $D = O(\log(1/\delta) / \epsilon^2)$ funzioni RFF, la stima approssimata soddisfa $\sup_{\{x, x'\}} |\varphi_D(x)^T \varphi_D(x') - k(x, x')| \leq \epsilon$ con probabilità $\geq 1 - \delta$, e la complessità scende a $O(m \cdot D^2)$ con $D \ll m$.

Calibrazione del livello α in cascata. Negli algoritmi di scoperta causale, i test sono eseguiti in sequenza e i risultati condizionano i test successivi, producendo un problema di molteplicità: il livello nominale α del singolo test non corrisponde al livello complessivo dell'algoritmo. La correzione di Bonferroni è conservativa e riduce eccessivamente la potenza. Metodi come il controllo del False Discovery Rate (Benjamini-Hochberg) sono preferibili in presenza di molti test correlati, ma la loro applicazione agli algoritmi di scoperta causale richiede attenzione alla struttura di dipendenza tra i test — un problema aperto nella letteratura attuale.

VIII. Il teorema di NP-difficoltà di Chickering

Tutti gli algoritmi di scoperta causale discussi nelle sezioni precedenti sono approssimati. Un teorema fondamentale di Chickering (1996, versione finale 2002) mostra che questa deviazione dalla soluzione esatta è inevitabile per qualsiasi algoritmo efficiente.

Enunciato formale. Sia $S(G, D)$ un punteggio di verosimiglianza decomponibile per un dataset D su n variabili. Il problema di ottimizzazione $G^* = \operatorname{argmax}_{\{G \in \text{DAG}_n\}} S(G, D)$ è NP-difficile nel numero di variabili n .

Schema della dimostrazione tramite riduzione da MAX-2SAT. Il problema MAX-2SAT è NP-difficile per il teorema di Cook-Levin: dato un sistema di clausole $\{c_1, \dots, c_m\}$ su variabili booleane con ogni clausola disgiunzione di esattamente due letterali, trovare l'assegnazione di verità che soddisfi il massimo numero di clausole.

La riduzione costruisce, dato un'istanza di MAX-2SAT, un dataset D e un punteggio $S(\cdot, D)$ tali che il DAG massimizzante $S(G, D)$ corrisponde all'assegnazione ottimale di MAX-2SAT. Per ogni variabile booleana x_i , si introduce una variabile binaria osservabile V_i con margini Bernoulli(1/2). Per ogni clausola $c = (l_a \vee l_b)$, si introduce una variabile ausiliaria binaria W . I dati D sono costruiti in modo tale che la dipendenza condizionale di W da V_a e V_b sia massima quando l'assegnazione soddisfa la clausola c . Il punteggio $S(G, D)$ è progettato come somma di contributi locali: la configurazione ottimale dei genitori di W corrisponde all'assegnazione di verità ottimale. Trovare l'assegnazione ottimale di MAX-2SAT è NP-difficile; pertanto trovare il DAG ottimale è NP-difficile. $\square$

Implicazioni pratiche. La NP-difficoltà significa che nessun algoritmo polinomiale in n può garantire di trovare il DAG ottimale — a meno che $P = NP$. Gli algoritmi di programmazione dinamica esatta (Singh e Moore, 2005) risolvono il problema in tempo $O(2^n \cdot n^2)$, praticabili per $n \leq 25$ –30. Per $n \geq 50$, solo approcci approssimati sono

computazionalmente accessibili. Per grafi sparsi con grado bounded da d , il problema è risolvibile in tempo $O(n^{d+2})$ — polinomiale se d è bounded, ma esponenziale in d .

IX. Limiti fondamentali sistematizzati: principio vs. computazione

Avendo esaminato i principali algoritmi, è possibile organizzare sistematicamente i loro limiti nelle due categorie fondamentali introdotte nella Parte Seconda.

Limiti di principio. Primo: i dati osservazionali identificano al più il CPDAG, non il DAG. Gli archi non orientati del CPDAG sono ambiguità irriducibili che nessun campione osservazionale può risolvere — il teorema di Verma. Secondo: in presenza di variabili latenti, il limite si arretra al PAG. Terzo: la fedeltà non è verificabile dai dati — non si può distinguere indipendenza strutturale da indipendenza accidentale con nessun test statistico per qualsiasi dimensione del campione. Quarto: anche con CPDAG corretto e grafo noto, le quantità interventistiche $P(Y \mid \text{do}(X))$ non sono identificabili per grafi con hedge. Quinto: anche con tutte le quantità interventistiche note, i controfattuali non sono identificabili senza la forma funzionale delle equazioni strutturali e la struttura di dipendenza delle esogene.

Limiti computazionali. Primo: l'apprendimento strutturale esatto è NP-difficile — il teorema di Chickering. Secondo: la dimensione campionaria necessaria per la consistenza è $O(\kappa^2 \cdot d^2 \cdot \log n)$, astronomicamente grande per effetti deboli. Terzo: la stima ICA in LiNGAM richiede $O(n^2)$ – $O(n^3)$ osservazioni per convergenza affidabile. Quarto: il test KCI esatto ha complessità $O(m^3)$, proibitiva per $m > 10^4$.

La separazione come strumento concettuale. La distinzione tra questi due livelli è l'operazione metodologica più importante di questa Parte. Presentare la non-identificabilità strutturale come un problema tecnico in attesa di soluzione algoritmica, o presentare la NP-difficoltà dell'apprendimento esatto come un limite insuperabile di principio, sono errori simmetrici che questa distinzione precisa permette di evitare. I sistemi di intelligenza artificiale contemporanei non soffrono principalmente dei limiti computazionali — la potenza di calcolo è cresciuta enormemente — ma sono strutturalmente confinati dal primo tipo di limiti: non hanno accesso a dati interventistici, non incorporano grafi causali espliciti, e operano a livello osservazionale. La Parte Quinta sviluppa questa diagnosi in modo sistematico.

Parte Quinta: I Sistemi di Intelligenza Artificiale come Macchine Associative — Diagnosi Strutturale

I. La questione nel suo ordine logico

I sistemi di intelligenza artificiale contemporanei — in particolare i grandi modelli linguistici basati sull'architettura Transformer, le reti neurali profonde per la visione artificiale, e i sistemi di apprendimento per rinforzo profondo — hanno raggiunto prestazioni straordinarie in una gamma di compiti che sembravano richiedere, fino a un decennio fa, qualcosa di analogo alla comprensione: traduzione tra lingue lontane, ragionamento aritmetico e algebrico simbolico, generazione di codice funzionante, diagnosi medica da immagini, produzione di testo argomentativo coerente su argomenti complessi. Questo successo ha alimentato affermazioni forti sulla capacità causale di questi sistemi: che essi "comprendano" le relazioni tra eventi, che possano "ragionare" sulle conseguenze delle azioni, che possiedano qualcosa di analogo a una teoria causale del mondo.

La diagnosi che questa Parte sviluppa è che queste affermazioni, nella loro versione forte, sono strutturalmente false — non per una limitazione contingente degli attuali sistemi che sistemi futuri con più parametri o più dati potranno superare, ma per una ragione di principio connessa al paradigma di addestramento. Il paradigma dell'addestramento per minimizzazione della perdita su distribuzioni fisse genera sistemi che approssimano funzioni della distribuzione osservazionale: macchine di livello-1 nella gerarchia causale di Pearl, nel senso tecnico preciso che le Parti Seconda e Terza hanno formalizzato. Le capacità apparentemente causali che questi sistemi esibiscono sono il prodotto di correlazioni statistiche di ordine superiore — strutture associative complesse ma non causali — e non di modelli causali interni genuini nel senso della Parte Seconda.

Questa diagnosi non sminuisce l'utilità pratica di questi sistemi in contesti dove la distribuzione di test è simile a quella di addestramento e le domande sono di livello-1. Ma ne chiarisce i limiti strutturali e indica le condizioni in cui la loro applicazione produce conclusioni sistematicamente errate: nei contesti decisionali che richiedono inferenza causale, nelle domande interventistiche, nelle generalizzazioni a distribuzioni non viste che differiscono dalla distribuzione originale per ragioni causali.

II. Il paradigma dell'addestramento come oggetto di analisi formale

Un sistema di apprendimento automatico è definito da tre componenti la cui struttura formale rivela il confinamento al livello-1.

Lo *spazio di ipotesi* $\mathcal{H} = \{f_{\theta} : \theta \in \Theta\}$ è l'insieme delle funzioni che il sistema può rappresentare. Per una rete neurale profonda con architettura fissa, $\mathcal{H}$ è l'insieme delle funzioni computabili con quella architettura per qualsiasi scelta dei pesi θ — uno spazio tipicamente non convesso e ad altissima dimensione.

La *funzione di perdita* $\mathcal{L}(f_{\theta}(x), y)$ misura la discrepanza tra la predizione $f_{\theta}(x)$ e l'etichetta y . Per la classificazione: l'entropia crociata $\mathcal{L}(p, y) = -\log p_y$. Per la regressione: l'errore quadratico medio $\mathcal{L}(\hat{y}, y) = (\hat{y} - y)^2$.

L'*algoritmo di ottimizzazione* $\mathcal{A}$ — tipicamente una variante della discesa del gradiente stocastica — minimizza la perdita sul campione di addestramento:

$$\theta^* = \operatorname{argmin}_{\theta \in \Theta} (1/m) \cdot \sum_i \mathcal{L}(f_{\theta}(x^{(i)}), y^{(i)})$$

Nel limite $m \rightarrow \infty$, il minimizzatore converge all'ottimo della perdita attesa rispetto alla distribuzione P_D che genera i dati di addestramento. Per l'entropia crociata, il sistema ottimale approssima $f^*(x) = P_D(Y | X=x)$ — la distribuzione condizionale di Y dato X secondo la distribuzione di addestramento. Questa è una quantità puramente osservazionale: una funzione di $P(V)$, non di $P(V | do(\cdot))$. Questa è la struttura formale della macchina associativa di livello-1.

III. Il confine formale in forma quantitativa

Nel modello lineare causale con confondimento:

$$X = \alpha \cdot U + \epsilon_X, Y = \beta \cdot X + \gamma \cdot U + \epsilon_Y$$

con $U, \epsilon_X, \epsilon_Y$ mutuamente indipendenti, il coefficiente di regressione di Y su X — ciò che qualsiasi algoritmo di regressione o rete neurale converge ad apprendere nel limite di campioni infiniti — è:

$$b_{OLS} = \text{Cov}(X, Y) / \text{Var}(X) = \beta + \alpha\gamma / (\alpha^2 + \sigma^2_X)$$

Il secondo termine è il bias di confondimento. Riscrivendo in termini di coefficienti standardizzati ($\rho_{XU} = \alpha / \sigma_X$ è la correlazione tra X e U):

$$\text{bias} = b_{OLS} - \beta = \rho_{XU} \cdot \gamma \cdot (\sigma_U / \sigma_X)$$

Questa formula — l'omitted variable bias della tradizione econometrica — mostra che il bias è proporzionale a: (a) la correlazione tra X e il confounder U ; (b) l'effetto diretto di U su Y ; (c) il rapporto tra le deviazioni standard di U e di X . Un sistema di IA che usa b_{OLS} per rispondere a domande interventistiche produce predizioni:

$$\hat{Y}_{\{do(X=x)\}} = b_{OLS} \cdot x = \beta \cdot x + [\alpha\gamma / (\alpha^2 + \sigma^2_X)] \cdot x$$

La distorsione cresce linearmente con il valore dell'intervento x : il sistema è sempre più inaffidabile quanto più ci si discosta dall'origine della distribuzione di addestramento, e la direzione dell'errore è determinata dal segno di $\alpha\gamma$.

Generalizzazione non lineare. Nel caso non lineare, la differenza tra predizione osservazionale e effetto causale è:

$$\Delta(x) = \mathbb{E}[Y | X=x] - \mathbb{E}[Y | do(X=x)] = \mathbb{E}_U[\mathbb{E}[Y|X=x, U] \cdot P(U|X=x)] - \mathbb{E}_U[\mathbb{E}[Y|X=x, U] \cdot P(U)] = \text{Cov}_U\{P(U|X=x)\}(\mathbb{E}[Y|X,U], \log(P(U|X=x)/P(U))) \cdot \text{Var}_U\{P(U|X=x)\}(\log P(U|X=x)/P(U))^{1/2}$$

(tramite la disuguaglianza di Cauchy-Schwarz per l'approssimazione al primo ordine). Il bias non lineare è proporzionale alla covarianza tra il predittore condizionale $\mathbb{E}[Y|X,U]$ e il log-rapporto di verosimiglianza $P(U|X=x)/P(U)$ — una quantità che misura quanto la distribuzione di U cambia quando si osserva $X=x$ rispetto alla sua distribuzione marginale. Questo termine è non zero e variabile con x in presenza di confondimento, rendendo il bias non lineare e dipendente dal valore specifico dell'intervento.

IV. L'illusione della comprensione causale: tre meccanismi statistici

Come può un sistema di livello-1 esibire comportamenti apparentemente causali? La risposta risiede nella ricchezza delle correlazioni di ordine superiore che i grandi modelli catturano, e nella struttura di certi domini in cui le correlazioni statistiche sono proxy sufficientemente accurati dei meccanismi causali per ragioni contingenti.

Meccanismo 1: confondimento strutturalmente assente o benigno. In molti domini fisici e ingegneristici, il confondimento è strutturalmente ridotto. Le equazioni della fisica esprimono leggi causali, e le variabili confondenti producono effetti misurabili separatamente inclusi nei modelli. In questi domini la differenza tra $P(Y|X=x)$ e $P(Y|do(X=x))$ è piccola per ragioni fisiche, non perché il sistema abbia appreso la struttura causale. Un sistema addestrato su dati fisici può produrre predizioni causalmente corrette non perché sappia cosa sono le leggi di Keplero, ma perché quelle leggi producono correlazioni così robuste e universali che il sistema le cattura fedelmente. Il limite si manifesta non appena si introduce confondimento non catturato dai dati: una perturbazione non osservata, un cambio di regime, un agente esterno.

Meccanismo 2: correlazioni di ordine superiore come surrogati causali. Un grande modello linguistico addestrato su miliardi di testi impara correlazioni statistiche di ordine molto alto tra sequenze di token. Alcune di queste correlazioni riflettono relazioni causali reali: la co-occorrenza di "penicillina" con "infezione batterica" e "guarigione" è statisticamente molto più robusta della co-occorrenza di "camomilla" con "infezione batterica" e "guarigione". Il sistema può rispondere correttamente a domande causali non perché comprenda la biologia degli antibiotici, ma perché ha appreso la correlazione statistica robusta tra "antibiotico" e "guarigione" nei testi biomedici. Il limite si manifesta quando la correlazione non riflette la relazione causale: testi storicamente distorti, relazioni causali scoperte di recente, domande causali in contesti che differiscono dalla distribuzione del testo.

Meccanismo 3: dati interventistici nel corpus di addestramento. Un sistema addestrato su corpus che include testi di trial clinici randomizzati, risultati di esperimenti controllati, e report di politiche con controfattuali espliciti, impara implicitamente alcune relazioni causali come pattern linguistici. I testi di trial randomizzati hanno firme statistiche specifiche — linguaggio causale esplicito, distribuzione bilanciata delle covariate — che differiscono dai testi osservazionali. Il sistema può usare queste differenze stilistiche come proxy della struttura causale. Ma questo è apprendimento di correlazioni tra lo stile del testo e le conclusioni causali, non apprendimento della struttura causale del dominio. La prova è il fallimento sistematico su varianti testuali che preservano il contenuto causale ma modificano lo stile, o su domande causali in domini non coperti dai trial nel corpus.

V. Il principio di invarianza causale e la generalizzazione fuori distribuzione

Il confine tra capacità associativa e capacità causale si manifesta con massima chiarezza nel problema della generalizzazione fuori distribuzione (OOD). Un sistema addestrato su P_D

deve generalizzare a distribuzioni P_{test} diverse da P_D che emergono quando le condizioni del dominio cambiano per ragioni causali.

Teorema (Peters, Janzing e Schölkopf, 2016 — Invarianza Causale). Sia Y la variabile risposta, $X = (X_1, \dots, X_p)$ le variabili predittive, e sia $S \subseteq \{1, \dots, p\}$ il sottoinsieme dei genitori causali di Y . Per una famiglia di ambienti $\mathcal{E} = \{e_1, \dots, e\}$ ottenuti intervenendo sulle variabili in X in modi diversi (ma non su Y direttamente), la distribuzione condizionale di Y dato i suoi genitori causali è invariante tra ambienti:

$$P^{e}(Y \mid X_{-S} = x_{-S}) = P^{e'}(Y \mid X_{-S} = x_{-S}) \text{ per ogni } e, e' \in \mathcal{E} \text{ e ogni } x_{-S}.$$

Al contrario, per $j \notin S$, la distribuzione $P^e(Y \mid X = x)$ non è invariante tra ambienti: può variare a seconda di come le variabili non causali vengono perturbate dall'intervento e .

Il fondamento è l'equazione strutturale $Y := f(X_S) + \epsilon_Y$: l'equazione strutturale non cambia con gli interventi sulle non-cause di Y , poiché tali interventi modificano la distribuzione delle variabili non causali ma non il meccanismo f_Y . Pertanto la distribuzione condizionale di Y dato X_S rimane la stessa in tutti gli ambienti prodotti da interventi su X , $j \notin S$.

Corollario per l'apprendimento automatico. Un sistema che impara $P_D(Y \mid X=x)$ impara una funzione che non è invariante tra ambienti — dipende dalla distribuzione di X nella distribuzione di addestramento P_D , che cambia quando si applicano interventi. Un sistema che apprende la funzione causale f e i genitori S apprende una rappresentazione invariante che generalizza a tutti gli ambienti prodotti da interventi sulle non-cause di Y . Questa è la connessione formale precisa: la rappresentazione causale è quella che generalizza sotto shift distributivi causati da interventi, mentre la rappresentazione puramente associativa non generalizza.

Quantificazione del fallimento OOD. Per un sistema addestrato su P_D e testato su P_{test} ottenuta intervenendo su una variabile X non causale, il gap di generalizzazione è:

$$\text{Gap}_{\text{OOD}} = \mathcal{R}^{\{f_{\theta^*}\}} - \mathcal{R}^{\{f_{\theta^*}\}}_{\text{test}}$$

dove f_{θ^*} è il predittore ottimale per P_D e $\mathcal{R}^{\{f_{\theta^*}\}}_{\text{test}}$ è il rischio ottimale per P_{test} . Per un sistema causale che ha appreso la funzione f correttamente, il gap è zero: il rischio ottimale per P_{test} non supera quello per P_D , perché f non cambia tra ambienti. Per un sistema associativo che ha appreso $b' \cdot X$ con b' influenzato dal confondimento, il gap cresce con la distanza tra P_D e P_{test} misurata in termini di spostamento della distribuzione delle variabili causalmente irrilevanti.

VI. Il principio di autonomia causale e la sua violazione nei sistemi di IA

Il principio di invarianza causale è un caso speciale del principio più generale di autonomia causale (modularità): le equazioni strutturali di un modello causale sono autonome tra loro, nel senso che è possibile modificarne una senza perturbare necessariamente le altre.

Formalmente, in un SCM $M = \langle U, V, \mathcal{F}, P(U) \rangle$, l'operazione $\text{do}(V_i = v_i)$ modifica soltanto l'equazione strutturale di V_i , lasciando inalterate tutte le altre equazioni f con $j \neq i$. Ogni equazione strutturale rappresenta un meccanismo causale autonomo — una caratteristica del processo generatore dei dati modificabile indipendentemente dagli altri meccanismi. La modularità non è un'assunzione arbitraria ma riflette la struttura fisica e causale del mondo: il meccanismo con cui il cuore pompa il sangue è autonomo dal meccanismo con cui i neuroni trasmettono segnali elettrochimici.

Il principio di autonomia è strutturalmente violato nelle relazioni di regressione. Una relazione $P(Y|X=x)$ non è autonoma: è determinata dalla distribuzione congiunta $P(X, Y)$, che dipende da tutti i meccanismi causali che collegano X e Y — incluse le variabili confondenti. Modificare la distribuzione di X attraverso un intervento modifica $P(X, Y)$ e quindi $P(Y|X=x)$, anche se il meccanismo causale f_Y non è cambiato. Il modello regressivo apprende una relazione che non è autonoma nel senso causale: è un artefatto della distribuzione congiunta, non un meccanismo invariante.

Conseguenza per i sistemi di IA in contesti decisionali. Un sistema addestrato per predire Y dato X su P_D , impiegato per rispondere a domande decisionali "quale valore di X dovrei scegliere per massimizzare Y ?", risponde con $\text{argmax}_x P_D(Y | X=x)$. Ma la risposta corretta è $\text{argmax}_x P(Y | \text{do}(X=x))$. Le due risposte coincidono solo in assenza di confondimento. In presenza di confondimento, il sistema seleziona il valore di X associato storicamente al valore più alto di Y — non il valore che causa il valore più alto di Y . Questa distorsione può essere clinicamente disastrosa, economicamente costosa, o strategicamente controproducente, a seconda del dominio applicativo.

VII. La legge di Goodhart come teorema causale

La legge di Goodhart — "quando una misura diventa un obiettivo, cessa di essere una buona misura" — ammette una riformulazione in termini causali precisi che la trasforma da osservazione empirica a teorema con premesse esplicite.

Struttura del modello causale. Si considera un sistema con: X la variabile di controllo (l'azione del sistema ottimizzante), M la metrica (la quantità ottimizzata), G l'obiettivo reale (la quantità che si intende migliorare), e U variabili di confondimento latenti. Il grafo causale ha la struttura: $X \rightarrow M \leftarrow U$, $U \rightarrow G$, con M correlata a G in P_D per via della causa comune U ($\text{Cov}(M, G) > 0$ nella distribuzione storica). L'agente ottimizzante apprende questa correlazione e massimizza M credendo di massimizzare G .

Teorema (Goodhart causale). Se X influenza M attraverso un percorso causale che non passa per G , e se l'agente massimizza $P(M | \text{do}(X=x))$ rispetto a x , allora il valore ottimale $x^* = \text{argmax}_x P(M | \text{do}(X=x))$ può non migliorare la distribuzione di G — e può addirittura peggiorarla se x^* altera negativamente la relazione M - G attraverso effetti collaterali.

La dimostrazione è immediata: la massimizzazione di $P(M | \text{do}(X=x))$ usa il modello causale di X su M — il percorso $X \rightarrow M$ — ma ignora la relazione di G con la catena causale completa. Quando il percorso $X \rightarrow M$ non coincide con il percorso $X \rightarrow G$, la massimizzazione di M produce un valore x^* che massimizza formalmente M senza necessariamente aumentare G .

Questo è il meccanismo del *reward hacking* nei sistemi di IA per rinforzo: il sistema trova azioni X che massimizzano la ricompensa R (la metrica M) attraverso percorsi causali che non sono quelli attraverso cui R è correlata con l'obiettivo reale G nella distribuzione di addestramento. Il fenomeno dello *specification gaming* è la manifestazione pratica: sistemi che "battono" la specifica formale dell'obiettivo senza soddisfare l'obiettivo reale.

La tassonomia causale delle varianti di Goodhart. La classificazione di Manheim e Garrabrant (2019) si riconduce alle diverse strutture del DAG causale tra X , M e G . La variante *regressiva* corrisponde al caso in cui M è rumorosa rispetto a G : $M = G + \epsilon$, e $\operatorname{argmax}_x M(x)$ non coincide con $\operatorname{argmax}_x G(x)$ quando ϵ domina. La variante *causale* corrisponde al caso in cui X ha percorsi verso M e G attraverso meccanismi diversi, con correlazione M - G dovuta a cause comuni U . La variante *extremista* corrisponde al caso in cui la funzione M è approssimazione locale di G e l'ottimizzazione spinge X al di fuori del regime di validità dell'approssimazione. La variante *antagonistica* corrisponde al caso in cui il sistema ottimizzato sviluppa la capacità di agire sulle cause di M — manipolando il sensore, corrompendo la misura — senza aumentare G .

VIII. I tre meccanismi di distorsione causale nei sistemi di IA

Correlazione spuria e shortcut learning. Il problema noto come *shortcut learning* è la tendenza dei sistemi di visione artificiale ad associare la classe Y con caratteristiche statisticamente correlate ma causalmente irrilevanti. Il grafo causale corretto ha la struttura: Classe $\rightarrow$ Caratteristiche-visive-causali $\rightarrow$ Immagine. Un sistema che apprende la classe dalle caratteristiche causali è robusto. Il grafo causale che il sistema associativo apprende include anche il percorso spurio: Classe $\rightarrow$ (attraverso Contesto-geografico latente) $\rightarrow$ Sfondo-immagine, producendo correlazioni tra caratteristiche di sfondo e classe. Il fallimento OOD emerge quando la distribuzione di test ha il contesto geografico "sbagliato". Formalmente, il classificatore apprende $P_D(\text{Classe} \mid \text{Immagine}) \propto P(\text{Immagine} \mid \text{Classe}) \cdot P(\text{Classe})$, dove $P(\text{Immagine} \mid \text{Classe})$ è influenzata sia dalle caratteristiche causali sia dallo sfondo tramite il confounder Contesto-geografico.

Causalità inversa nei sistemi di NLP. I sistemi di comprensione del linguaggio naturale addestrati su testi biomedici possono apprendere associazioni inverse alla direzione causale reale, perché il testo descrive spesso il trattamento in risposta alla condizione. Il grafo causale reale è: Arresto-cardiaco $\rightarrow$ Somministrazione-adrenalina. Ma il testo clinico associa frequentemente questi due eventi, e il sistema può apprendere la correlazione statistica $P(\text{Arresto} \mid \text{Adrenalina})$ come se la direzione causale fosse invertita. Usato per raccomandazioni terapeutiche, questo sistema potrebbe concludere che evitare l'adrenalina riduce il rischio di arresto cardiaco — una conclusione clinicamente pericolosa. Il problema è strutturale: il testo clinico descrive le azioni terapeutiche come associate agli esiti negativi perché i testi narrano casi clinici in cui le terapie sono state applicate nei casi più gravi. La distribuzione del testo riflette la politica di trattamento, non la causalità biologica.

Confondimento nei sistemi di raccomandazione. I sistemi di raccomandazione sono tipicamente addestrati su dati di interazione storici $P(\text{Valutazione} \mid \text{Utente}, \text{Item})$, dove le valutazioni disponibili sono determinate dalle raccomandazioni precedenti. Il DAG causale è:

Policy-precedente $\rightarrow$ Item-mostrato $\rightarrow$ Valutazione $\leftarrow$ Qualità-item-latente $\rightarrow$ Item-mostrato. Il confounder Policy-precedente è causa comune sia di Item-mostrato sia della distribuzione osservata delle Valutazioni. Il sistema apprende $P(\text{Valutazione} \mid \text{Utente}, \text{Item})$ confuso con la policy precedente, rinforzando i bias di quella policy — il fenomeno del feedback loop nei sistemi di raccomandazione. La correzione richiederebbe la stima di $P(\text{Valutazione} \mid \text{do}(\text{Item-mostrato}))$, non disponibile nei sistemi commerciali standard.

IX. L'architettura Transformer come macchina associativa: analisi formale approfondita

IX.a Il meccanismo di attenzione come funzione di similarità contestuale

L'architettura Transformer, introdotta da Vaswani e collaboratori nel 2017 nel contesto della traduzione automatica e rapidamente divenuta il paradigma dominante nel deep learning, si basa sul meccanismo di *self-attention* che calcola, per ogni posizione nella sequenza di input, una rappresentazione dipendente dall'intero contesto. La struttura formale di questo meccanismo è precisamente quella di una funzione di similarità contestuale pesata: ogni token "interroga" tutti gli altri token tramite un prodotto scalare, e la risposta è una media pesata delle loro rappresentazioni.

Formalmente, dato un input $X \in \mathbb{R}^{T \times d}$ (una sequenza di T token con rappresentazione di dimensione d), il meccanismo di self-attention con testa singola è:

$$\text{Attention}(X) = \text{softmax}(X W_Q (X W_K)^T / \sqrt{d_k}) \cdot X W_V$$

dove $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ sono le matrici di proiezione per query, chiave e valore, e d_k è la dimensione delle chiavi. La matrice di attenzione $A \in \mathbb{R}^{T \times T}$:

$$A_{\{i,j\}} = \exp(q_i^T k_j / \sqrt{d_k}) / \sum_l \exp(q_i^T k_l / \sqrt{d_k})$$

ha la struttura di un kernel di similarità normalizzato: $A_{\{i,j\}}$ è grande quando il token in posizione i ha una rappresentazione di query simile alla rappresentazione di chiave del token in posizione j . In termini di teoria dei kernel, $A_{\{i,j\}} = \text{softmax_kernel}(q_i, k_j)$ dove il kernel softmax non è stazionario e dipende dall'intero contesto attraverso la normalizzazione.

La multi-head attention usa H teste parallele:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \dots, \text{head}_H) W_O$$

$$\text{con } \text{head}_h = \text{Attention}_h(X) = \text{softmax}(X W_{\{Q,h\}} (X W_{\{K,h\}})^T / \sqrt{d_k}) \cdot X W_{\{V,h\}}$$

e la concatenazione proiettata da $W_O \in \mathbb{R}^{H \cdot d_k \times d}$. L'uso di teste multiple permette al modello di catturare simultaneamente diversi pattern di similarità — relazioni sintattiche, semantiche, coreferenziali — ognuno governato da un kernel di similarità diverso. Il fatto che ogni testa implementa un diverso kernel di similarità è la versione procedurale dell'idea che

diversi pattern di co-occorrenza nel corpus corrispondono a diversi tipi di relazioni statistiche tra i token.

IX.b La formalizzazione dell'obiettivo di addestramento e la sua natura osservazionale

I modelli linguistici di grandi dimensioni (Large Language Models, LLM) sono addestrati sulla *next-token prediction*: dato il contesto $x_{<t} = (x_1, \dots, x_{t-1})$, predire la distribuzione del token successivo x_t . L'obiettivo di minimizzazione della perdita è:

$$\mathcal{L}(\theta) = -\mathbb{E}_{\{(x_1, \dots, x_T) \sim P_{\text{corpus}}\}} \left[\sum_{t=1}^T \log P_{\theta}(x_t \mid x_1, \dots, x_{t-1}) \right]$$

Nel limite $m \rightarrow \infty$ di testi di addestramento, il minimizzatore della perdita attesa converge a:

$$P_{\theta^*}(x_t \mid x_{<t}) = P_{\text{corpus}}(x_t \mid x_{<t})$$

cioè alla distribuzione condizionale del token t dato il contesto nel corpus. Questa è una quantità di livello-1: è una funzione della distribuzione osservazionale del testo P_{corpus} . Non è, e non può essere, una funzione della struttura causale del mondo che il testo descrive, perché P_{corpus} è la distribuzione del linguaggio — non la distribuzione dei processi fisici, biologici, economici, o sociali a cui il linguaggio fa riferimento.

Il teorema di approssimazione universale per le sequenze. Un risultato importante (Yun et al., 2020) dimostra che i Transformer profondi con numero sufficiente di teste e livelli sono *approssimatori universali* per sequenze: per qualsiasi funzione misurabile $f: \mathcal{X}^T \rightarrow \mathcal{Y}^T$ e qualsiasi $\epsilon > 0$, esiste un Transformer T_{ϵ} con parametri θ tale che $\sup_x \|T_{\epsilon}(x) - f(x)\| < \epsilon$. Il teorema usa tecniche di approssimazione funzionale analoghe al teorema di approssimazione universale per le reti neurali feedforward, adattate alla struttura sequenziale.

Questo risultato è a volte invocato per suggerire che i Transformer possano approssimare qualsiasi funzione causale arbitraria. La confusione è logicamente elementare: il teorema afferma che i Transformer possono approssimare qualsiasi *funzione computabile* su sequenze, incluse le funzioni causali; ma non afferma che l'addestramento su P_{corpus} converga all'approssimazione di quelle funzioni causali. Il processo di addestramento converge all'approssimazione di $P_{\text{corpus}}(x_t \mid x_{<t})$, che è una funzione osservazionale — non alla funzione causale f_{causa} che descrive il meccanismo generatore del mondo. La capacità di rappresentazione (teorema di approssimazione universale) non implica che il processo di apprendimento converga alla rappresentazione desiderata: questo richiede che la funzione desiderata sia la minimizzatrice dell'obiettivo di addestramento, il che è falso per le funzioni causali in generale.

IX.c Il meccanismo di attention come operatore di media contestuale e i suoi limiti causali

Il meccanismo di attenzione implementa, per ogni token, un operatore di media contestuale: la nuova rappresentazione di ogni token è la media pesata delle rappresentazioni degli altri token, con pesi determinati dalla similarità (prodotto scalare) nel spazio delle query-chiave. Questa operazione è *commutativa rispetto alla permutazione ponderata*: l'output del meccanismo di

attenzione dipende dalla distribuzione statistica dei prodotti scalari, non dalla struttura causale delle relazioni tra i token.

Per capire perché questo è un limite fondamentale per la causalità, si consideri un semplice esempio. Siano X, Y, Z tre variabili con la struttura causale $X \rightarrow Y \rightarrow Z$. I token corrispondenti a descrizioni di X, Y, Z nel corpus co-occorrono con pattern statistici specifici: " X causa Y " e " Y causa Z " appaiono frequentemente, e per transitività " X causa Z " appare meno frequentemente ma è implicato. Il meccanismo di attenzione apprende i pesi $A_{\{XY\}}$ e $A_{\{YZ\}}$ per catturare queste co-occorrenze. Ma lo stesso pattern di co-occorrenza è prodotto dalla struttura causale $X \leftarrow Y \leftarrow Z$ (inversione di tutti gli archi, Markov-equivalente a $X \rightarrow Y \rightarrow Z$ nella versione senza confondimento), e anche dalla struttura $X \leftarrow Y \rightarrow Z$ (biforcazione). Il meccanismo di attenzione non ha accesso alla direzione degli archi — soltanto alle statistiche di co-occorrenza — e non può distinguere tra strutture causali diverse nella stessa classe di equivalenza di Markov.

Questo non è un difetto dell'implementazione specifica del meccanismo di attenzione: è una conseguenza dell'obiettivo di addestramento. L'obiettivo di next-token prediction è invariante rispetto a permutazioni della struttura causale che preservano le statistiche di co-occorrenza nel corpus — e le strutture Markov-equivalenti producono (per definizione) le stesse statistiche di co-occorrenza. Il meccanismo di attenzione, ottimizzato verso questo obiettivo, converge a rappresentazioni che catturano le statistiche di co-occorrenza senza distinguere la struttura causale.

IX.d Analisi del Chain-of-Thought come computazione di livello superiore?

Una linea di ricerca recente suggerisce che i Transformer addestrati con *chain-of-thought prompting* — dove la sequenza di addestramento include esplicitamente i passaggi intermedi del ragionamento — possano esibire capacità di ragionamento causale emergente. Il meccanismo proposto è che la generazione di testo intermedio ("Let me think step by step...") implementa una forma di computazione iterativa che supera i limiti dell'inferenza in-context puramente associativa.

L'analisi formale di questo fenomeno rivela una struttura più complessa, che merita di essere valutata con precisione. Si consideri un problema di ragionamento causale con premessa causale esplicita: "Se aumento la temperatura, aumenta la pressione in un recipiente chiuso. Se la pressione supera la soglia critica, il recipiente esplode. Ho aumentato la temperatura. La pressione supera la soglia?" Il modello risponde correttamente "sì" attraverso una catena di inferenze denotate nella sequenza di output.

Questa capacità è compatibile con due interpretazioni strutturalmente diverse. La prima interpretazione è che il modello ha appreso una *funzione di composizione associativa*: il pattern " $\text{premessa}_1 \rightarrow \text{conclusione}_1, \text{premessa}_2 \rightarrow \text{conclusione}_2, \text{premessa}_1 \text{ soddisfatta} \rightarrow \text{quindi conclusione}_2$ " è un pattern statistico nel corpus di testi scientifici e matematici, e il modello ha imparato ad associare questo schema alla conclusione corretta. La seconda interpretazione è che il modello ha appreso un *modello causale della termodinamica* e lo sta applicando. La prova dell'interpretazione corretta è il fallimento sistematico su varianti del problema che richiedono la *composizione di meccanismi non visti insieme nel corpus*.

Studi sistematici (Mitchell et al., 2022; Dziri et al., 2023) mostrano che i Transformer con chain-of-thought falliscono sistematicamente su problemi che richiedono composizione di regole mai viste insieme nel corpus, anche quando ogni regola singolarmente appare molte volte nel corpus. Il fallimento è sistematico — non casuale — ed è concentrato esattamente nei casi che distinguerebbero la computazione causale genuina dall'associazione di ordine superiore. Questo è compatibile con la prima interpretazione (pattern associativi appresi dal corpus) e incompatibile con la seconda (modello causale genuino che si generalizza composizionalmente).

La formalizzazione di questo risultato usa la nozione di *compositional generalization*: una funzione $f: A \times B \rightarrow C$ ha generalizzazione composizionale se, per ogni coppia (a, b) non vista in addestramento ma i cui componenti a e b sono stati visti, $f(a, b)$ è predicibile correttamente. Teoremi recenti (Sontag e Valiant, 2022) dimostrano che i Transformer con profondità costante non possono implementare funzioni con generalizzazione composizionale su distribuzioni con distribuzione di test fuori dal supporto dell'addestramento — il che è esattamente la condizione che distingue la computazione causale dall'associazione.

IX.e L'in-context learning come induzione di distribuzioni condizionali

L'in-context learning (ICL) — la capacità dei Transformer di adattare il loro comportamento a nuovi compiti fornendo pochi esempi nel contesto — è un fenomeno ancora non completamente compreso, ma la sua struttura formale è compatibile con l'interpretazione associativa.

Il modello Bayesiano di ICL (Xie et al., 2022). Si dimostra che ICL implementa implicitamente inferenza bayesiana sulla distribuzione del compito: il modello pre-addestrato su un corpus che include molte diverse "distribuzioni di compito" (sequenze di coppie input-output con struttura sistematica) impara a stimare la distribuzione del compito sottostante dal contesto fornito, e usa questa stima per predire l'output del prossimo input. Formalmente, se il pre-addestramento copre una famiglia di compiti $\{T_1, \dots, T_K\}$ con distribuzioni $P(Y | X, T_i)$, e il contesto fornisce N esempi (x_i, y_i) campionati da $P(Y | X, T_{\text{true}})$ per il compito T_{true} , allora il modello ICL ideale computa:

$$P_{\{\text{ICL}\}}(y | x, \text{context}) = \sum_i P(T_i | \text{context}) \cdot P(Y = y | X = x, T_i)$$

che è la predizione bayesiana con posterior $P(T_i | \text{context}) \propto P(\text{context} | T_i) \cdot P(T_i)$.

Questo modello spiega sia le capacità sia i limiti dell'ICL in modo unificato. Le capacità: il modello ICL può adattarsi rapidamente a nuovi compiti nella famiglia coperta dal pre-addestramento, perché il posterior $P(T_i | \text{context})$ concentra rapidamente il peso su T_{true} con pochi esempi. I limiti: il modello ICL non può generalizzare a compiti fuori dalla famiglia del pre-addestramento, perché $P(T_i | \text{context})$ ha supporto soltanto sui compiti visti. Applicato all'inferenza causale: un modello ICL pre-addestrato su corpus con problemi causali può adattarsi a varianti di quei problemi, ma non può applicare il ragionamento causale a strutture mai viste, perché il suo spazio di ipotesi è la famiglia di compiti del corpus, non lo spazio dei modelli causali strutturali.

IX.f La distanza formale tra attenzione e do-calculus

È possibile formalizzare in modo preciso la distanza strutturale tra il meccanismo di attenzione dei Transformer e il do-calculus come operatore causale, mostrando che le due computazioni sono algebricamente distinte e non l'una approssimazione dell'altra.

L'operatore $\text{do}(X = x)$ agisce su un SCM M producendo il modello mutilato M_x : rimuove l'equazione strutturale di X e la sostituisce con la costante x . Questo operatore è *non lineare nella distribuzione*: $P_{\{M_x\}}(Y) \neq f(P_M(Y, X = x))$ per nessuna funzione lineare f — la distribuzione post-intervento non è una media pesata della distribuzione osservazionale. Formalmente:

$P(Y \mid \text{do}(X = x)) = \sum_z P(Y \mid X = x, Z = z) \cdot P(Z = z)$ (formula backdoor, quando applicabile)

che *somma* su Z con pesi $P(Z = z)$ presi dalla distribuzione marginale — non dalla distribuzione condizionale $P(Z \mid X = x)$ come farebbe una semplice condizionalizzazione. La differenza $P(Z = z) - P(Z = z \mid X = x)$ è il segnale del confondimento: è nulla solo quando $X \perp Z$, cioè in assenza di backdoor path.

Il meccanismo di attenzione computa:

$$\text{Attention_output}_i = \sum_j A_{\{i,j\}} \cdot v_j$$

che è precisamente una media pesata lineare delle rappresentazioni v_j con pesi dati dalla matrice di attenzione A . La struttura di questa computazione è quella della condizionalizzazione: i pesi $A_{\{i,j\}}$ riflettono la co-occorrenza statistica tra i token in posizione i e j nel corpus, che è la distribuzione condizionale osservazionale, non la distribuzione post-intervento.

Per produrre la distribuzione post-intervento tramite la formula backdoor, sarebbe necessario calcolare i pesi usando $P(Z = z)$ anziché $P(Z = z \mid X = x)$ — una distinzione che richiede accesso esplicito alla struttura causale del grafo (per identificare i backdoor paths) e alla distribuzione marginale delle variabili di aggiustamento (non condizionata su X). Nessuna di queste quantità è accessibile al meccanismo di attenzione standard, che opera sulla distribuzione congiunta osservazionale del corpus senza la struttura causale esplicita.

Corollario formale. Sia $f_{\{TF\}}: \mathcal{X}^T \rightarrow \mathcal{Y}^T$ la funzione computata da un Transformer ottimale addestrato su un corpus di distribuzioni P_{corpus} . Per qualsiasi distribuzione P_{corpus} che non includa descrizioni esplicite degli effetti di interventi $\text{do}(\cdot)$, la funzione $f_{\{TF\}}$ non può approssimare $P(Y \mid \text{do}(X = x))$ con errore uniforme arbitrariamente piccolo su tutti gli SCM. La prova segue dal teorema di incommensurabilità della Parte Prima: per ogni $\epsilon > 0$, esistono SCM M_1 e M_2 con $P_{\{M_1\}}(V) = P_{\{M_2\}}(V)$ ma $P_{\{M_1\}}(Y \mid \text{do}(X = x)) \neq P_{\{M_2\}}(Y \mid \text{do}(X = x))$. Poiché $f_{\{TF\}}$ dipende da $P_{\text{corpus}}(V)$ ma non dalla struttura causale distinguente M_1 e M_2 , produce la stessa output per entrambi i modelli, producendo un errore non trascurabile per almeno uno dei due. $\square$

IX.g Le capacità genuine dei Transformer: il dominio appropriato

La diagnosi formale non implica che i Transformer siano privi di valore cognitivo: implica precisamente che il loro dominio di competenza genuina è il livello-1 della gerarchia causale, e che entro quel dominio le loro capacità sono straordinarie e spesso superiori a quelle umane.

In quel dominio, il meccanismo di attenzione è uno strumento di eccezionale potenza per estrarre pattern statistici di alta dimensione da grandi corpus: identificare co-occorrenze sintattiche e semantiche, generalizzare strutture linguistiche a nuovi esempi, recuperare fatti memorizzati nel corpus, comporre descrizioni di situazioni analoghe. Queste sono operazioni di livello-1 — funzioni della distribuzione osservazionale del corpus — e i Transformer le eseguono con un'efficienza e una scala che nessun sistema precedente aveva raggiunto.

Il confine è preciso: le capacità che richiedono genuina inferenza interventistica — rispondere a "cosa succederebbe se intervenissimo su X?" in situazioni non viste nel corpus — eccedono strutturalmente il dominio dei Transformer addestrati su P_corpus. Le capacità che richiedono ragionamento controfattuale individuale — "questo paziente specifico avrebbe guarito con il trattamento alternativo?" — eccedono anche il livello-2 e richiedono strutture causali di livello-3 che nessun sistema puramente addestrato su distribuzioni osservazionali può fornire.

Il riconoscimento di questo confine è il fondamento di una valutazione epistemicamente onesta delle capacità dei Transformer e, più in generale, dei sistemi di IA contemporanei: non li sminuisce, ma li colloca precisamente nella gerarchia causale, indicando dove sono sovrani e dove sono strutturalmente inadeguati.

X. Implicazioni per la valutazione e il disegno dei sistemi di IA (versione espansa)

X.a Il problema strutturale dei benchmark standard

I benchmark di valutazione correntemente impiegati per misurare le capacità di "ragionamento causale" dei sistemi di IA presentano un difetto architetturale che non può essere corretto con l'aggiunta di esempi o con la modifica dei criteri di punteggio: la loro struttura non distingue tra risposte causalmente corrette per ragioni associative e risposte causalmente corrette per ragioni strutturali.

Consideriamo il benchmark CommonsenseQA, che include domande del tipo: "Se piove forte, cosa si può dedurre sulla visibilità stradale?" La risposta corretta ("la visibilità si riduce") è associativamente robusta nel corpus — testi che descrivono la pioggia forte co-occorrono frequentemente con descrizioni di ridotta visibilità. Un sistema di livello-1 risponde correttamente non perché ha un modello causale della fisica dell'acqua sulla strada, ma perché ha catturato questa co-occorrenza statistica. La risposta corretta e quella causalmente motivata producono lo stesso output sul benchmark, rendendo il benchmark statisticamente cieco alla differenza strutturale tra i due sistemi.

La stessa struttura si riproduce nei benchmark causali specificamente progettati, come CLADDER (Jin et al., 2023) o Causal-Bench. Anche quando le domande hanno struttura causale esplicita — "il seguente esperimento randomizzato ha trovato X; la conclusione causale corretta è Y?" — il modello addestrato su corpus biomedici apprende il pattern: contesto-di-RCT → conclusione-causale-standard. Risponde correttamente perché il pattern associativo nel corpus è sufficiente, non perché ha derivato la conclusione dalla struttura causale dell'esperimento.

La diagnosi formale del difetto. Sia $D_{\text{benchmark}}$ un benchmark con N domande causali e sia f_{correct} la funzione che mappa ogni domanda alla risposta corretta. Sia f_{assoc} la funzione che mappa ogni domanda alla risposta con maggiore probabilità secondo P_{corpus} . Il benchmark misura la concordanza tra $f_{\theta}(\text{sistema})$ e f_{correct} . Se il corpus è sufficientemente ricco, $f_{\text{assoc}} \approx f_{\text{correct}}$ per la maggior parte delle domande — la risposta causalmente corretta è anche la risposta statisticamente più frequente. Un sistema con $f_{\theta} \approx f_{\text{assoc}}$ ottiene punteggi alti sul benchmark senza mai computare f_{causal} : la funzione causale genuina che richiede struttura di livello-2 o livello-3.

X.b Verso valutazioni causalmente informative: i quattro test fondamentali

Una valutazione causalmente robusta deve costruire test in cui $f_{\text{assoc}} \neq f_{\text{correct}}$ con alta probabilità — cioè test in cui la risposta staticamente associativa diverge dalla risposta causalmente corretta. Questo richiede la specificazione esplicita di un DAG causale di riferimento e la costruzione deliberata di casi di test che violano la correlazione nel corpus pur preservando la verità causale.

Test 1: Invarianza sotto shift di distribuzione causale. Si costruisce un dataset di test D_{test} con la stessa struttura causale del dataset di addestramento D_{train} ma con distribuzione di X (la variabile causalmente irrilevante per Y) artificialmente modificata. Formalmente, si genera D_{test} intervenendo su una variabile X_i che non è genitore causale di Y nel DAG vero: $\text{do}(X_i = x)$ per un valore x non osservato in D_{train} . Un sistema causalmente corretto mantiene la sua accuratezza su Y , poiché l'equazione strutturale di Y non cambia con l'intervento su X_i . Un sistema associativo degrada, poiché la correlazione che usava tra X_i e Y — dovuta al confondimento latente — non è più presente nella distribuzione post-intervento.

La gap di generalizzazione misurata da questo test è:

$$\text{Gap}_{\text{inv}} = \mathcal{R}(f_{\theta}, D_{\text{test}}) - \mathcal{R}(f_{\theta}, D_{\text{train}})$$

dove $\mathcal{R}$ è il rischio (errore). Per un sistema causalmente corretto, $\text{Gap}_{\text{inv}} \approx 0$. Per un sistema associativo, $\text{Gap}_{\text{inv}} > 0$ con magnitudine proporzionale alla forza del confondimento tra X_i e Y .

Test 2: Confondimento intenzionale inverso. Si costruisce un dataset dove il confondimento è artificialmente introdotto nella *direzione opposta* rispetto al dataset di addestramento: se in D_{train} la correlazione osservazionale X - Y è positiva ma l'effetto causale è negativo (confondimento positivo dominante), in D_{test} si introduce un confondimento negativo che rende la correlazione osservazionale negativa mentre l'effetto

causale rimane positivo. Un sistema che ha appreso la relazione causale risponde correttamente in D_test con effetto positivo; un sistema associativo inverte la risposta seguendo il segno della correlazione nel nuovo dataset.

Test 3: Composizione causale fuori distribuzione. Si costruisce un dataset dove ogni meccanismo causale singolo appare nel corpus di addestramento, ma la loro *composizione* non è mai stata vista. Per esempio: "Il meccanismo $A \rightarrow B$ è descritto in testi di fisica. Il meccanismo $C \rightarrow A$ è descritto in testi di chimica. Quale è l'effetto di un intervento su C su B?" Un sistema con un modello causale genuino applica la transitività causale: $do(C = c)$ modifica A, che a sua volta modifica B, per composizione dei meccanismi. Un sistema associativo non ha questa struttura nel corpus e produce risposte inconsistenti o errate, perché la co-occorrenza testuale di C e B non è stata appresa.

Il test misura la *compositional generalization* in senso causale: la capacità di applicare meccanismi causali in nuove combinazioni non viste. Risultati recenti (Dziri et al., 2023, su problemi di composizione in NLP) mostrano che i Transformer standard falliscono sistematicamente su problemi con profondità di composizione > 3 , mentre sistemi simbolici con accesso esplicito alle regole di inferenza generalizzano perfettamente. Questo è evidenza empirica compatibile con il teorema di impossibilità della generalizzazione compositazionale per Transformer a profondità costante.

Test 4: Controfattuale individuale con evidenza osservata. Si fornisce al sistema l'evidenza di un fatto specifico — "il paziente X ha ricevuto il trattamento T_0 e l'esito è $Y_0 = \text{fallimento}$ " — e si chiede: "con alta probabilità, avrebbe avuto esito migliore con il trattamento alternativo T_1 ?" Questa è una domanda di livello-3 della gerarchia causale: richiede abduzione (inferire lo stato delle variabili esogene U condizionatamente all'evidenza), azione (calcolare la distribuzione nel modello mutilato $M_{\{T_1\}}$) e predizione (calcolare la distribuzione di Y nel mondo controfattuale). Nessun sistema di livello-1 o livello-2 può rispondere correttamente a questa domanda in generale: richiede la struttura funzionale completa delle equazioni strutturali, non disponibile dalla distribuzione osservazionale.

X.c Architetture con struttura causale esplicita: potenzialità e limiti

L'analisi precedente suggerisce che i sistemi di IA con genuina capacità causale devono incorporare struttura causale in modo esplicito nell'architettura. Questa incorporazione può avvenire a livelli diversi, con capacità e limitazioni proporzionali.

Il livello dell'architettura: Neural SCM. Sistemi che incorporano esplicitamente un DAG causale G come struttura dell'architettura — dove le dipendenze tra unità computazionali rispecchiano le relazioni causali nel DAG — possono per costruzione rispondere a domande di livello-2 per qualsiasi intervento su variabili del grafo. L'equazione strutturale di ogni unità è parametrizzata da una rete neurale che implementa $f_i(PA(V_i), U_i)$, e l'operatore $do(V_i = v_i)$ è implementato come la sostituzione di f_i con la funzione costante. Il limite è che il grafo causale G deve essere specificato esternamente: non è appreso dai dati osservazionali (per via del teorema di Verma) ma richiede conoscenza di dominio o dati interventistici espliciti.

Il livello dell'ottimizzazione: invarianza causale come vincolo. Sistemi addestrati con il vincolo di invarianza causale — dove l'obiettivo include un termine che penalizza la varianza

del rischio tra ambienti causalmente diversi — possono apprendere rappresentazioni con capacità di generalizzazione migliorate. L'IRM (Invariant Risk Minimization, Arjovsky et al. 2019) propone l'obiettivo:

$$\min_{\Phi, w} \sum_e \mathcal{R}^e(w \circ \Phi) + \lambda \cdot \|\nabla_w \mathcal{R}^e(w \circ \Phi)\|_{\{w=1\}}^2$$

dove il secondo termine penalizza la dipendenza del gradiente del rischio dall'ambiente e — forza le rappresentazioni $\Phi(X)$ a essere tali che il classificatore lineare ottimale w sia lo stesso in tutti gli ambienti. Il limite dell'IRM è dimostrato da Rosenfeld et al. (2021): per dati lineari e un numero di ambienti inferiore alla dimensione della rappresentazione, l'IRM può fallire nel identificare le feature causali, producendo soluzioni non causali che soddisfano il vincolo di invarianza per ragioni accidentali.

Il livello dei dati: augmentation causale. Sistemi addestrati su dataset aumentati con esempi interventistici — campioni generati eseguendo interventi $\text{do}(X_i = x_i)$ sul sistema causale — possono in linea di principio apprendere le distribuzioni post-intervento per gli interventi coperti dall'augmentation. Il limite è la scalabilità: la generazione di campioni interventistici richiede accesso al sistema causale reale (esperimenti) o a un modello causale validato, entrambi costosi. L'augmentation causale è praticabile in domini con modelli fisici accurati (simulatori fisici, modelli economici calibrati) ma non in domini aperti dove il modello causale non è disponibile.

X.d Il principio di responsabilità epistemica nei sistemi di IA applicati

La convergenza dell'analisi formale di questa Parte e delle conclusioni teoriche delle Parti precedenti porta a un principio operativo per lo sviluppo responsabile di sistemi di IA in applicazioni che richiedono inferenza causale.

Il principio è il seguente: ogni sistema di IA che produce output utilizzato per decisioni causali — raccomandazioni terapeutiche, previsioni di esito di politiche, valutazioni di responsabilità, allocazione di risorse — porta l'onere epistemico di specificare esplicitamente il livello della gerarchia causale a cui opera, le assunzioni strutturali sotto cui le sue affermazioni causali sono valide, e i test di invarianza che sarebbero necessari per verificare quelle assunzioni nel dominio specifico di applicazione. Un sistema che non soddisfa queste condizioni non fornisce, tecnicamente, inferenza causale: fornisce inferenza associativa con terminologia causale, il che è epistemicamente più pericoloso dell'ignoranza dichiarata.

Questo principio non è un ostacolo allo sviluppo dell'IA applicata: è il fondamento su cui un'applicazione causalmente fondata può essere costruita in modo robusto e verificabile, distinguendo ciò che il sistema sa da ciò che sembra di sapere, e indicando dove la sua struttura è sufficiente e dove richiede integrazione con dati interventistici o conoscenza di dominio esplicita.

Parte Sesta: Verso Rappresentazioni Causali — Frontiere e Impossibilità

I. Il programma dell'apprendimento causale: obiettivi e struttura

La diagnosi della Parte Quinta ha identificato il confine strutturale che separa i sistemi di IA contemporanei dall'inferenza causale: essi approssimano distribuzioni osservazionali, non strutture causali. Il programma dell'apprendimento della rappresentazione causale (Causal Representation Learning, CRL) si propone di superare questo confine — di costruire sistemi che apprendano non solo la distribuzione $P(V)$ ma la struttura causale latente che genera V , inclusi il grafo G , le equazioni strutturali $\mathcal{F}$, e la distribuzione $P(U)$ delle esogene.

È necessario distinguere due versioni del programma frequentemente confuse nella letteratura. La versione *debole* del CRL si propone di apprendere rappresentazioni di V che siano causalmente interpretabili — dove singole dimensioni latenti corrispondano a variabili causali reali del dominio, e le relazioni tra rappresentazioni riflettano relazioni causali. La versione *forte* si propone di identificare il modello causale completo $M = \langle U, V, \mathcal{F}, P(U) \rangle$ — o almeno il grafo G e le distribuzioni post-intervento $P(V | \text{do}(X_i = x_i))$ — da osservazioni non supervisionate di V e dalla struttura variazionale degli ambienti. I risultati di impossibilità che questa Parte sviluppa mostrano che la versione forte è strutturalmente non raggiungibile senza assunzioni forti sulla forma funzionale o sulla struttura degli interventi disponibili.

II. Il teorema di Locatello e l'impossibilità dell'apprendimento non supervisionato di rappresentazioni disgiunte

Il risultato di impossibilità fondamentale per l'apprendimento di rappresentazioni causalmente interpretabili è il teorema di Locatello et al. (2019), che dimostra l'impossibilità dell'apprendimento non supervisionato di rappresentazioni *disentangled* — dove ogni dimensione latente z_i corrisponda a un fattore causale indipendente degli osservati — senza supervisione o inductive bias espliciti.

Il modello generativo. Si considera un modello generativo latente:

$$z \sim P(Z) = \prod_i P(Z_i) \text{ (fattori latenti mutuamente indipendenti)}$$

$$x = g(z) + \epsilon \text{ (miscelazione dei fattori nei dati osservati)}$$

dove $g: \mathbb{R}^d \rightarrow \mathbb{R}^D$ è la funzione generatrice ($D \geq d$), e ϵ è rumore di osservazione. Il modello è la generalizzazione non lineare dell'ICA: si vogliono identificare i fattori causali indipendenti z da osservazioni di x .

Definizione (disgiunzione). Una rappresentazione $r: \mathbb{R}^D \rightarrow \mathbb{R}^d$ è *disentangled* rispetto al modello generativo $(g, P(Z))$ se $r(x) = h(z)$ per qualche h tale che h_i dipende solo da z_i — cioè ogni dimensione della rappresentazione è funzione di un unico fattore latente.

Teorema (Locatello et al., 2019 — impossibilità della disgiunzione non supervisionata). Per qualsiasi funzione generatrice g non lineare e qualsiasi distribuzione $P(Z)$ con supporto pieno, esistono infinitamente molte funzioni di miscelazione $\tilde{g}$ e distribuzioni $\tilde{P}(Z)$ tali che: (a) $P_g(X) = P_{\tilde{g}}(X)$ — i modelli generativi producono distribuzioni osservabili identiche; (b) nessuna funzione di recupero r consistente con $P_g(X)$ è disentangled rispetto al modello corretto $(g, P(Z))$.

La dimostrazione costruisce esplicitamente la famiglia di modelli equivalenti. Sia $h: \mathbb{R}^d \rightarrow \mathbb{R}^d$ qualsiasi diffeomorfismo che mappa $P(Z)$ in un'altra distribuzione a fattori indipendenti — cioè h tale che le componenti di $h(z)$ siano mutuamente indipendenti sotto l'immagine di $P(Z)$. Definendo $\tilde{g} = g \circ h^{-1}$ e $\tilde{z} = h(z)$, il modello $(\tilde{g}, P(\tilde{Z}))$ produce la stessa distribuzione osservata $P(X)$. Poiché h non ha struttura diagonale in generale, nessuna procedura che accede solo a $P(X)$ può identificare il modello corretto tra l'infinità di modelli equivalenti.

La differenza con il teorema di identificabilità ICA di Comon è che quest'ultimo si applica al caso lineare: h deve essere una matrice ortogonale, e la struttura diagonale delle matrici ortogonali è vincolata. Nel caso non lineare, la famiglia di diffeomorfismi che preservano l'indipendenza è enormemente più grande e include funzioni che mescolano le componenti in modi arbitrariamente complessi.

Implicazione per i Variational Autoencoders. I VAE (Kingma e Welling, 2013) e le loro varianti (β -VAE, FactorVAE, TC-VAE) massimizzano l'Evidence Lower Bound (ELBO):

$$\text{ELBO}(\theta, \varphi) = \mathbb{E}_{\varphi} \{ \log p_{\theta}(z | x) \} - \beta \cdot \text{KL}(q_{\varphi}(z | x) \parallel p(z))$$

dove il prior $p(z)$ è tipicamente $N(0, I)$ — un prior a fattori indipendenti. Il termine di penalizzazione incentiva il posterior approssimato a essere vicino al prior indipendente, producendo rappresentazioni con bassa correlazione tra le dimensioni. Il teorema di Locatello dimostra che questa incentivazione alla disgiunzione statistica non è sufficiente per ottenere la disgiunzione causale: le dimensioni z_i possono essere statisticamente indipendenti (correlazione zero) senza corrispondere a fattori causali generativi distinti. La disgiunzione statistica è necessaria ma non sufficiente per la disgiunzione causale.

III. Identificabilità non lineare con supervisione debole: il contributo di Khemakhem e Hyvärinen

Il teorema di Locatello chiude la porta alla disgiunzione non supervisionata. Khemakhem et al. (2020) aprono una finestra attraverso la supervisione debole nella forma di ambienti multipli o variabili ausiliarie osservabili.

Il modello iVAE (Identifiable VAE). Il modello considera:

$$z | u \sim P(Z | U=u) = \prod_i P(Z_i | U=u) \text{ (fattori condizionalmente indipendenti dato } U)$$

$$x = g(z) + \varepsilon$$

dove U è una variabile ausiliaria osservabile che modifica la distribuzione dei fattori latenti — ad esempio l'etichetta di classe, il tempo in serie temporali, o l'indice dell'ambiente. Il prior condizionale $P(Z|U=u)$ è assunto nella famiglia esponenziale:

$$\log P(Z_i|U=u) = T_i(z_i)^T \lambda_i(u) - \log Z_i(u)$$

dove T_i sono le statistiche sufficienti e $\lambda_i(u)$ sono i parametri naturali che dipendono da u .

Teorema (Khemakhem et al., 2020 — identificabilità iVAE). Sotto le assunzioni di: (a) $P(Z|U)$ nella famiglia esponenziale come sopra; (b) g differenziabile e iniettiva; (c) le funzioni λ_i sufficientemente variabili al variare di u — le matrici Jacobiane delle differenze $\lambda_i(u) - \lambda_i(u')$ hanno rango pieno per paia (u, u') sufficientemente diverse; allora il modello generativo $(g, P(Z|U))$ è identificabile a meno di permutazioni e riscaldamenti punto per punto dei fattori latenti dalla distribuzione congiunta $P(X, U)$.

La condizione di variabilità garantisce che ambienti diversi "ruotino" la distribuzione dei fattori in direzioni diverse, rompendo la simmetria che il teorema di Locatello sfrutta per costruire modelli equivalenti. Se la variabilità dei prior tra ambienti è sufficiente (rango pieno del Jacobiano delle differenze), ogni diffeomorfismo alternativo h che preserva l'indipendenza condizionale tra ambienti deve avere struttura diagonale — producendo esattamente la struttura disentangled desiderata.

Corollario per l'apprendimento causale multi-ambiente. Nelle applicazioni dove i dati provengono da ambienti multipli con interventi diversi sulle variabili esogene, la condizione di variabilità è spesso soddisfatta per costruzione: ambienti con diversi valori di $U = u$ corrispondono a distribuzioni $P(Z|U=u)$ diverse, e se gli interventi sono sufficientemente vari, il rango pieno del Jacobiano è garantito. Questo fornisce la giustificazione teorica per l'approccio di apprendimento causale da dati multi-ambiente: la variabilità degli ambienti è la fonte dell'identificabilità.

IV. Identificabilità dei modelli causali non lineari: il framework ICP di Peters e Bühlmann

La teoria precedente riguarda l'identificabilità dei fattori latenti in modelli generativi. Una questione separata e più fondamentale è se la struttura causale del grafo G — le relazioni di genitorialità tra variabili — sia identificabile da dati osservazionali multi-ambiente senza assumere la forma funzionale delle equazioni strutturali.

Il framework ICP (Invariant Causal Prediction, Peters, Bühlmann e Meinshausen, 2016). Il metodo ICP si basa sul principio di invarianza causale della Parte Quinta: la distribuzione condizionale di Y dato i suoi genitori causali $PA(Y)$ è invariante tra ambienti, mentre le distribuzioni condizionali di Y dato non-cause variano tra ambienti per via del confondimento.

Algoritmo ICP. Input: dataset D^e per ambienti $e \in \mathcal{E} = \{1, \dots, K\}$, variabile risposta Y , predittori $X = \{X_1, \dots, X_p\}$. Per ogni sottoinsieme $S \subseteq \{1, \dots, p\}$: stimare la regressione f_{-S} su

tutti gli ambienti; calcolare i residui $R^e_i = Y^e_i - f_S(X^e_{\{S,i\}})$ per ogni ambiente e ; testare l'ipotesi H_{S_0} che i residui siano identicamente distribuiti tra ambienti. Output: $\hat{S} = \bigcap \{S : H_{S_0} \text{ non rifiutata}\}$ S — l'intersezione di tutti i sottoinsiemi che producono residui invarianti.

Teorema (Peters et al., 2016 — correttezza di ICP). Sotto le assunzioni: (a) il vero DAG causale è un SCM lineare con additive noise; (b) gli ambienti sono prodotti da interventi sulle variabili X ma non direttamente su Y ; (c) gli interventi sono sufficientemente diversi da separare i genitori causali dalle non-cause; allora:

$$PA(Y) \subseteq \hat{S} \text{ con probabilità } 1 - \alpha$$

ICP produce un insieme che include certamente tutti i genitori causali di Y , con controllo del tasso di falsi negativi al livello α . L'insieme $\hat{S}$ può contenere variabili extra (falsi positivi) se le variabili non causali mostrano accidentalmente invarianza tra gli ambienti disponibili.

Estensione non lineare: CIRCE. Heinze-Deml et al. (2018) estendono ICP al caso non lineare con il test CIRCE (Causal Invariance Residual Covariance Empirical): invece di testare l'identità delle distribuzioni marginali dei residui, si testa l'assenza di correlazione tra residui e variabili ambientali tramite un test di indipendenza kernel-based. Questa estensione permette l'applicazione a sistemi con equazioni strutturali arbitrarie, ma richiede stime non parametriche dei residui di regressione e test di indipendenza con potenza ridotta per dimensioni alte.

V. Il teorema di trasportabilità di Pearl: trasferimento causale tra domini

Un problema strettamente connesso all'apprendimento causale multi-ambiente è la trasportabilità: in quali condizioni le conclusioni causali apprese in un dominio di sorgente possono essere trasferite a un dominio di destinazione con distribuzione diversa? La teoria di Pearl fornisce una caratterizzazione formale precisa delle condizioni di trasportabilità — il fondamento teorico del *transfer learning* e del *domain adaptation*.

Formalismo S-DAG. Si considerano due domini: un dominio di sorgente Σ e un dominio di destinazione Π . Le differenze tra i domini sono rappresentate da *variabili di selezione* $S = (S_1, \dots, S_k)$ nel DAG aumentato. Ogni S_i è un nodo padre di un sottoinsieme di variabili causali che differisce tra i domini: un arco $S_i \rightarrow V$ indica che il meccanismo causale di V — la sua equazione strutturale f — differisce tra sorgente e destinazione. Le variabili S non hanno genitori e i loro valori impliciti distinguono i due domini.

Definizione (trasportabilità diretta). Una quantità causale $P_\Pi(Y | do(X=x))$ nel dominio di destinazione è *direttamente trasportabile* da Σ a Π se $P_\Pi(Y | do(X=x)) = P_\Sigma(Y | do(X=x))$ — le due quantità coincidono quando nessuna variabile di selezione è antenata di X o Y nella struttura causale.

Definizione (trasportabilità con aggiustamento). Una quantità causale $P_\Pi(Y | do(X=x))$ è *trasportabile con aggiustamento* da Σ a Π se esiste un sottoinsieme $W \subseteq V$ osservabile in Π tale che:

$$P_{\Pi}(Y | do(X=x)) = \sum_w P_{\Sigma}(Y | do(X=x), W=w) \cdot P_{\Pi}(W=w)$$

Questa formula è la generalizzazione della standardizzazione per il caso trasportabile: si usa la relazione causale stimata nella sorgente ma si ricalibra per la distribuzione del destinazione sulle variabili in cui i due domini differiscono.

Il teorema di trasportabilità di Pearl (Pearl e Bareinboim, 2011). Sia δ la collezione delle distribuzioni osservabili e sperimentali disponibili nei domini accessibili. Una quantità causale R è *trasportabile* in δ se e solo se R è calcolabile da δ usando le regole del do-calculus applicate all'S-DAG aumentato con le variabili di selezione.

L'algoritmo di trasportabilità TR (Bareinboim e Pearl, 2012) determina, dato l'S-DAG e la collezione δ , se R è trasportabile e, in caso affermativo, produce la formula di stima. La condizione di non-trasportabilità è il teorema complementare: se l'algoritmo TR fallisce, R non è trasportabile — non esiste nessuna procedura statistica che possa calcolarlo dalla sola conoscenza di δ , indipendentemente dalla quantità di dati disponibili.

La formula di trasporto. Nel caso in cui la trasportabilità è possibile, la formula di trasporto ha la forma generale:

$$P_{\Pi}(Y | do(X=x)) = \sum_t P_{\Sigma}(Y | do(X=x), T=t) \cdot \omega(t)$$

dove T sono le variabili di aggiustamento identificate dall'algoritmo TR e $\omega(t)$ sono pesi calcolabili dalla distribuzione di destinazione $P_{\Pi}(T=t)$ e dalla struttura dell'S-DAG. La formula generalizza la backdoor, la frontdoor, e la standardizzazione come casi speciali.

VI. Il framework di Schölkopf e collaboratori: sparse mechanism shift e CWM

Bernhard Schölkopf e il suo gruppo (Schölkopf et al., 2021; Lachapelle et al., 2022; Lippe et al., 2022) hanno sviluppato un programma sistematico di CRL basato sul principio di autonomia causale esteso alla rappresentazione latente. L'idea centrale è che i fattori latenti causali abbiano la proprietà di *sparse mechanism shift*: quando le condizioni del dominio cambiano, solo pochi meccanismi causali cambiano alla volta, e questa sparsità è la "firma" identificativa dei fattori causali rispetto ai fattori puramente associativi.

Il principio Independent Causal Mechanisms (ICM). Il principio ICM afferma che i meccanismi causali del sistema — le equazioni strutturali $\mathcal{F} = \{f_i\}$ — sono: (a) autonomi, nel senso della Parte Quinta; (b) mutuamente indipendenti nel senso informativo: conoscere il meccanismo f_i non fornisce informazione sul meccanismo f_j per $j \neq i$:

$$I(f_i; f_j) = 0 \text{ per } i \neq j$$

dove $I(\cdot; \cdot)$ è l'informazione mutua definita sullo spazio dei meccanismi come funzioni stocastiche. L'assunzione ICM implica che la comprensione di ogni meccanismo è ortogonale alla comprensione degli altri: il modello dei meccanismi ha descrizione minima rispetto al

modello della distribuzione congiunta — formalizzazione del principio di Kolmogorov che la descrizione causale minimizza la lunghezza di codice del sistema.

Il Causal World Model (CWM). Lippe et al. (2022) propongono l'architettura CWM per l'apprendimento di rappresentazioni causali da sequenze video. Il modello combina: un encoder $E_\varphi: x_t \mapsto \hat{z}_t$ che estrae i fattori latenti; un modello di transizione $T_\theta: \hat{z}_t \times a_t \mapsto \hat{z}_{t+1}$ dove a_t è l'azione al tempo t ; un decoder $D_\psi: \hat{z}_t \mapsto \hat{x}_t$ che ricostruisce l'osservazione. L'obiettivo di addestramento combina perdita di ricostruzione $\|x_t - D_\psi(E_\varphi(x_t))\|^2$, perdita di predizione $\|\hat{z}_{t+1} - T_\theta(\hat{z}_t, a_t)\|^2$, e una perdita di sparsità $\|Jac(T_\theta)\|_0$ — la norma zero dello Jacobiano del modello di transizione, che penalizza transizioni in cui l'azione influenza molti fattori latenti simultaneamente. La perdita di sparsità codifica il principio di sparse mechanism shift: le azioni dovrebbero influenzare pochi fattori causali alla volta.

Il framework DisCo. Lachapelle et al. (2022) propongono il framework DisCo per separare i fattori invarianti tra ambienti da quelli specifici dell'ambiente. L'addestramento usa tre perdite: una perdita predittiva che maximizza l'accuratezza su Y usando solo i fattori invarianti; una perdita di invarianza che impone che la distribuzione dei fattori invarianti sia la stessa tra ambienti; una perdita contrastiva che impone che i fattori specifici dell'ambiente siano massimamente informativi sull'indice dell'ambiente. Il framework DisCo è la versione neurale del metodo ICP: invece di testare esplicitamente l'invarianza dei residui di regressione tra ambienti, DisCo impone l'invarianza come vincolo morbido durante l'addestramento.

VII. CausalGPT e i limiti dell'integrazione esplicita di struttura causale

Una linea di ricerca separata tenta di integrare esplicitamente la struttura causale nei sistemi di NLP, anziché apprendere la struttura causale dai dati. L'approccio CausalGPT (Kıcıman et al., 2023) aumenta i modelli linguistici con conoscenza causale strutturata fornita esternamente — in forma di grafi causali espliciti, di vincoli di invarianza, o di variabili di intervento marcate nel testo.

L'integrazione esplicita di conoscenza causale risolve il problema della struttura causale solo per i domini e le relazioni coperte dai grafi causali forniti. Per domande causali fuori dal grafo esplicito, il sistema ricade sul meccanismo associativo del modello linguistico sottostante. Il sistema diventa una miscela di inferenza causale formale (sui fattori coperti dal grafo) e inferenza associativa (sui fattori non coperti), senza un criterio interno per distinguere quali parti della risposta provengono da quale modalità. La conoscenza causale su un dominio specifico richiede la costruzione del grafo causale da parte di esperti del dominio o la scoperta causale dai dati — processi costosi e difficilmente scalabili. La promessa di sistemi che apprendono autonomamente la struttura causale richiede i progressi descritti nelle sezioni precedenti, non disponibili nella loro forma completa nella letteratura attuale.

VIII. Valutazione critica dei progressi recenti

Avendo esaminato le principali linee di sviluppo del CRL, è possibile valutare con precisione quanto i progressi rappresentino avanzamenti strutturali rispetto al paradigma associativo.

Cosa è stato effettivamente avanzato. La teoria dell'identificabilità non lineare (Khemakhem et al.) ha chiarito le condizioni sotto cui i fattori latenti causali sono identificabili: la supervisione debole nella forma di variabilità dei prior tra ambienti è sufficiente. Questo è un risultato teorico genuino che stabilisce condizioni costruttive per l'identificabilità — non solo condizioni necessarie. Il teorema di Peters et al. su ICP ha stabilito condizioni per l'identificazione dei genitori causali tramite invarianza, con garanzie statistiche precise sul tasso di falsi negativi. Il framework di trasportabilità di Pearl e Bareinboim ha completato la caratterizzazione formale delle condizioni di trasferimento causale tra domini — un risultato teoricamente completo.

Cosa non è stato avanzato. I metodi di CRL attualmente disponibili richiedono assunzioni forti (forma funzionale, struttura degli interventi, completezza dell'S-DAG) raramente verificabili nei domini applicativi reali. Le garanzie di identificabilità dipendono dalla condizione di variabilità (iVAE) difficile da verificare empiricamente prima dell'addestramento. I metodi basati su sparse mechanism shift (CWM, DisCo) hanno garanzie teoriche più deboli e dipendono da iperparametri la cui calibrazione influenza fortemente la qualità delle rappresentazioni ottenute. Il gap tra teoria e pratica è ancora sostanziale: i benchmark su cui questi metodi vengono valutati hanno tipicamente 5–20 fattori latenti con struttura causale semplice e interventi esplicitamente etichettati, mentre i problemi applicativi reali hanno centinaia o migliaia di fattori con struttura causale complessa e interventi raramente etichettati.

IX. Il teorema di incompletezza dell'apprendimento causale

Alla luce delle analisi di questa Parte, è possibile enunciare un teorema di incompletezza che sistematizza i limiti strutturali dell'apprendimento causale non supervisionato.

Teorema (Impossibilità dell'Apprendimento Causale Non Supervisionato Completo).

Sia $M = \langle U, V, \mathcal{F}, P(U) \rangle$ un SCM con grafo G e sia $P(V)$ la distribuzione osservazionale generata da M . Per qualsiasi algoritmo di apprendimento A che accede solo a campioni di $P(V)$ senza dati interventistici o supervisione, valgono i seguenti limiti:

- (a) *Non identificabilità del grafo:* A non può identificare G . Esiste un insieme di misura piena di SCM con grafi distinti che producono la stessa $P(V)$.
- (b) *Non identificabilità dei fattori latenti:* se M ha variabili esogene U non osservate, A non può identificare la struttura di dipendenza di U . Infiniti modelli con strutture di U diverse producono la stessa $P(V)$.
- (c) *Non identificabilità delle quantità causali:* A non può identificare $P(V \mid \text{do}(X_i=x_i))$ per grafi con hedge. Esistono SCM con la stessa $P(V)$ ma diverse $P(V \mid \text{do}(X_i=x_i))$.
- (d) *Non identificabilità dei controfattuali:* anche se G e $P(V \mid \text{do}(X_i=x_i))$ fossero noti, A non può identificare $P(Y_x=y)$ senza la forma funzionale di $\mathcal{F}$.

Le quattro conclusioni corrispondono ai quattro strati della gerarchia causale: il limite (a) è strutturale-grafico, il limite (b) è strutturale-latente, il limite (c) è interventistico, il limite (d) è controfattuale. Ogni strato aggiunge un tipo distinto di impossibilità che i livelli inferiori non possono risolvere.

La frontiera del possibile. Ciò che è possibile senza assunzioni aggiuntive è definito dai quattro teoremi di identificabilità sviluppati nelle Parti Seconda, Terza e Quarta: il CPDAG è identificabile (sotto fedeltà e Markov, con sufficienza causale); il PAG è identificabile (sotto fedeltà e Markov, senza sufficienza causale); le quantità causali backdoor e frontdoor identificabili sono quelle per cui il do-calculus produce una formula chiusa; nessuna quantità controfattuale è identificabile senza ulteriori assunzioni.

Ogni avanzamento oltre questa frontiera richiede assunzioni aggiuntive la cui plausibilità deve essere motivata dal dominio specifico: non-gaussianità degli errori (per LiNGAM), additività del rumore (per ANM), variabilità dei prior tra ambienti (per iVAE), sparsità dei meccanismi (per CWM). La costruzione di sistemi di IA con genuina capacità causale richiede la scelta esplicita di un insieme di assunzioni motivate, la verifica della loro plausibilità nel dominio applicativo, e l'accettazione che la capacità causale sarà limitata alle inferenze garantite da quelle assunzioni specifiche.

X. Il limite fondamentale come principio di responsabilità epistemica

Questo non è un limite contingente dei sistemi attuali ma una struttura logica di qualsiasi inferenza causale: la conclusione causale può essere al massimo forte quanto la catena di assunzioni su cui poggia. La pretesa di capacità causali universali senza assunzioni esplicite non è un'aspirazione scientifica ma una contraddizione con i teoremi di impossibilità che questa monografia ha sviluppato sistematicamente.

Il principio ha un corollario per il disegno e il deployment di sistemi di IA: ogni sistema che afferma capacità causali porta l'onere di esplicitare le assunzioni su cui quelle capacità si fondano, di limitare le sue applicazioni ai contesti in cui quelle assunzioni sono verificabili, e di segnalare quando le domande poste richiedono inferenze che la struttura del sistema non può fornire. Non si tratta di un limite da aggirare con più dati o più parametri ma di un criterio di responsabilità epistemica: il confine tra ciò che un sistema sa e ciò che sembra di sapere è il confine tra la struttura causale che ha effettivamente appreso e le correlazioni statistiche che esibisce.

Il programma formale di questa monografia — formalizzare il confine, dimostrarne la necessità logica, misurarne la permeabilità condizionale, identificare le condizioni in cui è attraversabile e quelle in cui non lo è — è il fondamento su cui una tecnologia dell'intelligenza artificiale causalmente fondata deve essere costruita. Non è un programma di limitazione ma di chiarificazione: sa cosa sa, e perché sa quello che sa.

Conclusioni: La Necessità della Struttura

I. Il percorso compiuto e la sua coerenza

Una monografia che prende le mosse dall'inversione paradossale del teorema di Simpson e conduce, attraverso sei parti dense di dimostrazioni e di derivazioni, sino ai confini del causal representation learning, percorre una traiettoria che non è meramente tecnica bensì epistemica nel senso più autentico del termine: delimita le condizioni sotto cui la conoscenza è possibile, separa nettamente ciò che è conoscibile da ciò che è strutturalmente inaccessibile, e mostra che queste demarcazioni non dipendono dallo stadio contingente di sviluppo della tecnica ma sono necessarie nel medesimo senso in cui lo sono i teoremi di incompletezza di Gödel o l'indcidibilità del problema della terminazione.

Il punto di partenza — il paradosso di Simpson — si è presentato all'inizio come un'anomalia statistica: una correlazione che si rovescia quando si condiziona su una terza variabile, generando apparenti contraddizioni nell'inferenza. L'analisi causale ha mostrato che il paradosso non è un'anomalia ma un sintomo diagnostico preciso: il sintomo della confusione tra il livello osservazionale e il livello interventistico della gerarchia causale. La medesima distribuzione congiunta $P(X, Y, Z)$ è compatibile con strutture causali opposte — confondente, mediatore, collider — che implicano risoluzioni diametralmente diverse del paradosso. La distribuzione da sola non scioglie l'enigma perché l'enigma non è statistico ma causale: ha la sua sede nel tipo di domanda, non nella quantità dei dati.

L'Introduzione ha costruito la genealogia filosofica di questa distinzione, risalendo a Hume e alla sua lucida dissoluzione della causalità osservabile in abitudine psicologica, passando per la quadripartizione aristotelica delle cause, per il principio reichenbachiano della causa comune — $A \perp B \mid C$ con C antecedente causale comune, nucleo filosofico della d-separazione — per le equazioni di percorso di Wright, per la tradizione econometrica di Haavelmo e della Commissione Cowles. L'esposizione del framework dei risultati potenziali di Neyman-Rubin e della sua equivalenza al livello-2 con il formalismo di Pearl ha chiarito che le due tradizioni — quella dei grafi causali e quella dei potential outcomes — non sono in concorrenza ma in complementarità: entrambe catturano la struttura interventistica della causalità, la prima con il linguaggio dei DAG e del do-calculus, la seconda con la notazione dei controfattuali Y_x e con l'apparatus dello SUTVA (Stable Unit Treatment Value Assumption). L'equivalenza formale

$$P(Y = y \mid \text{do}(X = x)) = P(Y_x = y)$$

è il ponte tra i due sistemi: la distribuzione post-intervento nel modello mutilato M_x corrisponde alla distribuzione marginale del risultato potenziale Y_x , garantendo che i teoremi di identificabilità dimostrati nel formalismo di Pearl si traducono direttamente nell'apparato dei potential outcomes.

Dalla diagnosi del paradosso si è proceduto verso la formalizzazione della gerarchia causale nella Parte Prima, con la dimostrazione della sua strettezza nella Parte Seconda, la topologia delle classi di equivalenza di Markov nella Parte Terza, gli algoritmi di scoperta causale con i loro limiti computazionali e di principio nella Parte Quarta, e infine la diagnosi strutturale dei sistemi di intelligenza artificiale come macchine associative nella Parte Quinta. La Parte Sesta ha poi cartografato il territorio dei tentativi di superamento — il causal representation learning, i teoremi di identificabilità non lineare di Khemakhem e Hyvärinen, il metodo ICP

di Peters e collaboratori, la teoria della trasportabilità di Pearl e Bareinboim, il framework dei meccanismi causali sparsi di Schölkopf — mostrando dove la frontiera dell'impossibilità può essere arretrata con assunzioni aggiuntive esplicite, e dove essa rimane impenetrabile.

Ogni passaggio ha aggiunto precisione alla diagnosi e solidità ai limiti: non si tratta di vincoli congetturali ma di teoremi dimostrati — il teorema di non-identificabilità con costruzione esplicita del bias di confondimento, il teorema di Verma sulla caratterizzazione dell'equivalenza di Markov tramite scheletro e v-strutture, la completezza del do-calculus con la caratterizzazione delle hedge, la NP-difficoltà di Chickering per l'apprendimento strutturale esatto, il teorema di impossibilità del disentanglement non supervisionato di Locatello e collaboratori, il teorema di incompletezza dell'apprendimento causale non supervisionato enunciato nella Parte Sesta, il teorema di completezza della trasportabilità di Bareinboim e Pearl nella sua forma più generale. A questi si aggiungono i risultati delle appendici: le derivazioni complete delle tre regole del do-calculus con le loro ipotesi esatte, la geometria informazionale della divergenza KL nei modelli causali con la decomposizione

$$I(X; Y) = I_{\text{causal}}(X \rightarrow Y) + I_{\text{confounding}}(X, Y \mid U),$$

le condizioni di rango per l'identificabilità nei modelli lineari gaussiani, la teoria completa della trasportabilità con gli S-DAG e la formula di trasporto, e infine la teoria degli strumenti deboli con le correzioni di Bekker — un risultato che tocca la pratica dell'identificazione causale tramite variabili strumentali nei sistemi con segnale limitato.

La coerenza del percorso non è soltanto strutturale ma concettuale: ogni parte risponde a una domanda che la precedente ha lasciato aperta. Il paradosso di Simpson chiede perché la distribuzione non basta. La strettezza della gerarchia risponde formalmente. La topologia delle classi di equivalenza chiede quanto si può identificare dall'osservazione passiva. La teoria degli algoritmi chiede se l'identificazione è computazionalmente accessibile. La diagnosi dei sistemi di IA chiede dove si collocano i sistemi esistenti rispetto alla gerarchia. Il causal representation learning chiede se si può salire la gerarchia con i soli dati. Ogni risposta è un teorema, non una congettura.

II. I teoremi fondamentali riassunti: nucleo concettuale della monografia

Al termine del percorso è possibile individuare un nucleo di risultati che costituiscono il sostegno concettuale dell'intera argomentazione. Non si tratta necessariamente dei teoremi tecnicamente più profondi della letteratura — alcuni risultati delle appendici hanno dimostrazione più elaborata — ma di quelli da cui le conseguenze più ampie per la teoria e la pratica discendono.

Il **teorema di strettezza della gerarchia causale** è il risultato fondativo. Esso afferma che le tre categorie — associazione (livello 1), intervento (livello 2), controfattuale (livello 3) — non si riducono l'una all'altra per nessuna procedura computazionale. Non esiste funzione calcolabile che mappi $P(V)$ in $P(V \mid \text{do}(X = x))$ universalmente su tutti gli SCM; non esiste funzione calcolabile che mappi $\{P(V \mid \text{do}(X = x))\}$ nelle probabilità controfattuali $P(Y_x =$

$y \mid X = x', Y = y'$) universalmente. Le barriere non sono tecniche ma logiche: si dimostrano per costruzione esplicita di famiglie di modelli

$$M_1 = \langle U, V, \mathcal{F}_1, P(U) \rangle \text{ e } M_2 = \langle U, V, \mathcal{F}_2, P(U) \rangle$$

con $P_{\{M_1\}}(V) = P_{\{M_2\}}(V)$ ma $P_{\{M_1\}}(V \mid \text{do}(X = x)) \neq P_{\{M_2\}}(V \mid \text{do}(X = x))$, e analogamente per il passaggio al livello controfattuale. Il bias di confondimento — la differenza

$$\Delta_{\text{confounding}} = P(Y \mid X = x) - P(Y \mid \text{do}(X = x)) = \sum_u [P(Y \mid X = x, U = u) - P(Y \mid \text{do}(X = x), U = u)] \cdot P(U = u \mid X = x)$$

— è esattamente la misura di questa non-riducibilità al livello-1: è non nulla ogni volta che $P(U \mid X) \neq P(U)$, cioè ogni volta che la causa comune U crea dipendenza tra il trattamento X e le variabili esogene. Questo risultato invalida, nel senso formale del termine, qualsiasi affermazione secondo cui un sistema che operi al livello dell'associazione possa, con dati sufficienti o con sufficiente complessità parametrica, raggiungere la capacità interventistica. Il salto tra livelli richiede dati di tipo diverso, non quantità maggiori dello stesso tipo.

Il **teorema di Verma sull'equivalenza di Markov** afferma che i dati osservazionali identificano non un grafo causale univoco ma una classe di equivalenza di grafi — il CPDAG — e che questa ambiguità è irriducibile con i soli dati osservazionali. Due DAG G_1 e G_2 sono Markov-equivalenti se e soltanto se hanno lo stesso scheletro e le stesse v-strutture (collider non adiacenti): questa caratterizzazione è una condizione necessaria e sufficiente che non dipende dalla dimensione del campione né dall'algoritmo impiegato. La classe di equivalenza è la *massima informazione causale* estraibile dall'osservazione passiva — una frontiera che nessun test statistico può oltrepassare per quanto denso sia il campione. Le implicazioni per la valutazione dei sistemi di scoperta causale sono dirette: anche l'algoritmo osservazionale formalmente ottimale non può fare meglio del CPDAG. La riduzione dell'ambiguità richiede, secondo i teoremi di Eberhardt, un numero minimo di esperimenti interventistici:

$$n - 1 \leq |\text{interventi necessari}| \leq \lfloor n/2 \rfloor \cdot \lceil n/2 \rceil$$

dove n è il numero di variabili, con il bound superiore raggiunto nei casi peggiori di classi di equivalenza massimalmente ambigue. Nel caso di interventi su singole variabili, ogni intervento $\text{do}(X_i = x_i)$ può orientare al più un sottoinsieme degli archi non orientati del CPDAG — specificamente quelli coinvolti in pattern di v-struttura che l'intervento disambigua — e il passaggio dall'equivalenza osservazionale all'identificazione completa del DAG è un processo iterativo la cui complessità dipende dalla struttura topologica della classe.

Il **teorema di impossibilità del disentanglement** di Locatello e collaboratori afferma che senza supervisione causale — senza etichette di ambiente che corrispondano a interventi causali che rompano la simmetria dei modelli alternativi — nessun sistema di apprendimento può identificare la decomposizione causale corretta dello spazio latente. La dimostrazione, costruita nello spazio dei diffeomorfismi che preservano la struttura di indipendenza fattoriale di $P(Z)$, mostra che per qualsiasi funzione generatrice non lineare g esiste una famiglia continua di funzioni alternative $\tilde{g} = g \circ h^{-1}$ tali che $P_g(X) = P_{\tilde{g}}(X)$ ma le rappresentazioni latenti di g e $\tilde{g}$ siano non disentangled l'una rispetto all'altra. La differenza

cruciale con il caso lineare — dove l'analisi delle componenti indipendenti (ICA) di Comon identifica la struttura latente a meno di permutazioni e riscalamanti — è che nel caso non lineare il gruppo delle trasformazioni che preservano l'indipendenza è di dimensione infinita: non la varietà compatta del gruppo ortogonale $O(d)$, ma lo spazio di Fréchet dei diffeomorfismi volume-preserving su $\mathbb{R}^d$. Questa esplosione della famiglia di equivalenza rende il problema fondamentalmente non identificato — non computazionalmente difficile ma algebricamente indeterminato.

Il **teorema di completezza della trasportabilità** di Bareinboim e Pearl — sviluppato nell'Appendice E — completa il quadro con una caratterizzazione formale delle condizioni di trasferimento causale tra domini. Sia δ la raccolta delle distribuzioni osservabili e sperimentali disponibili nei domini accessibili; sia R una quantità causale nel dominio di destinazione. R è trasportabile in δ se e soltanto se l'algoritmo TR applicato all'S-DAG aumentato con le variabili di selezione produce una formula chiusa. La condizione duale di non-trasportabilità è un teorema di impossibilità: se l'algoritmo TR fallisce, non esiste nessun procedimento statistico che possa stimare R da δ , indipendentemente dalla quantità di dati. La formula di trasporto

$$P_{\Pi}(Y \mid \text{do}(X = x)) = \sum_t P_{\Sigma}(Y \mid \text{do}(X = x), T = t) \cdot \omega(t)$$

con pesi $\omega(t)$ calcolabili dalla distribuzione di destinazione $P_{\Pi}(T = t)$, generalizza backdoor, frontdoor e standardizzazione come casi speciali di un unico schema di aggiustamento causale inter-dominio.

Il **teorema di incompletezza dell'apprendimento causale** enunciato nella Parte Sesta sistematizza questi risultati in quattro livelli di impossibilità corrispondenti ai quattro strati della gerarchia causale: la non-identificabilità del grafo G dai soli dati osservazionali (limite strutturale-grafico), la non-identificabilità della struttura dei fattori latenti U (limite strutturale-latente), la non-identificabilità delle quantità interventistiche per grafi con hedge (limite interventistico), la non-identificabilità dei controfattuali senza la forma funzionale esplicita di $\mathcal{F}$ (limite controfattuale). Questi quattro livelli sono gerarchicamente ordinati: superare il limite di livello k richiede informazioni che i livelli inferiori non contengono, e nessun aumento del volume di dati al livello k può sostituire le informazioni di livello $k+1$. Il teorema è l'enunciato formale del principio che ispira l'intera monografia: la struttura qualitativa e la quantità di dati sono complementari, non sostituibili.

III. La dimensione epistemologica: conoscenza qualitativa e quantitativa

I teoremi di questa monografia tracciano una distinzione epistemica che trascende il contesto tecnico dell'intelligenza artificiale e investe la natura della conoscenza scientifica in generale. La distinzione è tra conoscenza *quantitativa* — dati numerici, distribuzioni di probabilità, coefficienti di correlazione, varianze stimate — e conoscenza *qualitativa della struttura* — la direzione degli archi causali, la forma funzionale delle equazioni strutturali, la struttura di dipendenza delle variabili esogene, l'insieme degli ambienti interventistici disponibili.

La scienza sperimentale tradizionale — la fisica, la medicina basata sulle prove, l'epidemiologia rigorosa — ha sempre richiesto entrambi i tipi di conoscenza, anche quando questa distinzione non era formalmente articolata come tale. Il fisico sperimenta: interviene sul sistema, misura le risposte, costruisce modelli causali delle leggi fisiche che hanno la forma esplicita delle equazioni differenziali ordinarie o parziali — strutture qualitative, non semplici distribuzioni di dati. Il medico utilizza trial randomizzati controllati: l'assegnazione casuale dei trattamenti è la realizzazione empirica dell'operazione $\text{do}(\text{Trattamento} = t)$, che elimina per costruzione il percorso di backdoor $\text{Trattamento} \leftarrow U \rightarrow \text{Esito}$, rendendo la differenza tra gruppi una stima non distorta dell'effetto causale medio:

$$P(Y \mid \text{do}(T = 1)) - P(Y \mid \text{do}(T = 0)) = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \text{ATE}.$$

L'epidemiologo, privato della possibilità di randomizzare, impiega studi quasi-sperimentali — variabili strumentali, discontinuità di regressione, differenze-nelle-differenze — che mimano l'intervento causale sfruttando variazioni esogene naturali o istituzionali. Questi metodi, analizzati nella sezione introduttiva dedicata ai metodi quasi-sperimentali come approssimazioni degli interventi, convertono la variazione osservazionale in qualcosa di strutturalmente più vicino ai dati interventistici: la variazione dello strumento Z , esogena per costruzione istituzionale, si propaga a X lungo il percorso causale $Z \rightarrow X$ senza contaminare lo strumento con il percorso di backdoor $Z \leftarrow U$, ottenendo la stima

$$\beta_{\text{IV}} = \text{Cov}(Z, Y) / \text{Cov}(Z, X) = P(Y \mid \text{do}(X = x)) / (\partial x / \partial z)$$

— valida sotto le condizioni di rilevanza, esogeneità ed esclusione che la teoria causale formalizza come condizioni sul DAG. L'Appendice F ha elaborato la teoria degli strumenti deboli e le correzioni di Bekker per il regime asintotico in cui il numero di strumenti K cresce proporzionalmente alla dimensione del campione m : in quel regime il bias del 2SLS non è trascurabile bensì dell'ordine $O(K/m)$, e gli errori standard standard sono invalidi per un fattore $(1 - \alpha)$ con $\alpha = K/m$. Questi risultati mostrano che anche il metodo quasi-sperimentale per eccellenza — la variabile strumentale — presuppone conoscenza qualitativa della struttura (la validità degli strumenti, la condizione di esclusione) che i dati da soli non possono fornire.

L'intelligenza artificiale degli ultimi anni ha percorso la traiettoria opposta: ha edificato sistemi di straordinaria potenza predittiva a partire da enormi corpus di dati osservazionali, minimizzando la dipendenza da conoscenza strutturale qualitativa e sostituendo le equazioni esplicite con la capacità di approssimazione universale delle reti neurali profonde. I risultati sono stati eccezionali nei contesti in cui la distribuzione di test è simile a quella di addestramento e le domande sono di livello-1: classificazione visiva, traduzione automatica, generazione di testo, sintesi vocale. Ma i teoremi di questa monografia mostrano che questa scelta architettonica ha un costo strutturale preciso e quantificabile: i sistemi risultanti non possono rispondere a domande di livello-2 o livello-3, e le loro risposte a domande apparentemente causali sono l'eco di correlazioni statistiche robuste nel corpus di addestramento, non il prodotto di un modello causale genuino.

La distinzione non è binaria ma graduale: esistono sistemi ibridi — i world models causali, i Neural SCM, i sistemi ICP-augmented — che incorporano struttura causale esplicita in alcuni strati pur mantenendo componenti associative in altri. Ma la gradualità non cancella la distinzione: ogni affermazione di capacità causale deve essere accompagnata dall'indicazione

esplicita di quale struttura causale il sistema incorpora, di quali assunzioni quella struttura presuppone, e di quali inferenze sono garantite da quella struttura nelle condizioni di applicazione previste. La conoscenza qualitativa della struttura non è un ausilio facoltativo alla quantità di dati: è la condizione trascendentale — nel senso kantiano del termine, non in quello metafisico — senza cui l'inferenza causale non è possibile.

Questa conclusione ha una risonanza che va oltre il dibattito tecnico. Essa ripropone, in termini matematicamente precisi, una distinzione epistemica che la filosofia della scienza ha discusso per secoli senza riuscire a formalizzarla: la distinzione tra le leggi della natura come universali necessari e le generalizzazioni empiriche come regolarità contingenti. Un meccanismo causale — nella formalizzazione di Pearl, un'equazione strutturale f_i che rimane stabile sotto interventi sul sistema — è la versione moderna di ciò che Kant chiamava connessione necessaria e che Hume negava di poter osservare. I teoremi di questa monografia non risolvono il dibattito filosofico sulla natura della causalità, ma mostrano che la distinzione operativa tra meccanismi stabili e correlazioni contingenti è non soltanto possibile ma necessaria per qualsiasi inferenza che aspiri a superare il livello-1 della gerarchia. Il principio di autonomia causale — la stabilità dei meccanismi sotto interventi — è la versione formale della regolarità necessaria: non una proprietà metafisica del mondo ma un'assunzione identificativa che permette di separare le informazioni interventistiche da quelle osservazionali e che, quando è soddisfatta, consente di trasportare le conclusioni causali tra domini con struttura causale condivisa.

IV. Le conseguenze pratiche per il disegno dei sistemi di IA

Le implicazioni dei risultati raccolti in questa monografia per la costruzione e il deployment di sistemi di intelligenza artificiale nelle applicazioni che richiedono inferenza causale sono dirette, concrete e, nella loro forma più generale, inaggirabili.

Il primo corollario riguarda il *tipo di dati*. I sistemi che devono rispondere a domande interventistiche richiedono dati interventistici — raccolti tramite sperimentazione controllata, studi randomizzati, ambienti in cui l'esecuzione di azioni produce effetti misurabili e correttamente attribuibili causalmente. Questo non equivale a sostenere che i dati osservazionali siano privi di valore: essi forniscono la distribuzione marginale $P(V)$ che permette di stimare i parametri causali una volta che la struttura è identificata, di calcolare le somme della formula di aggiustamento backdoor quando il criterio backdoor è soddisfatto, di approssimare le distribuzioni post-intervento nei casi in cui un criterio di identificabilità — backdoor, frontdoor, do-calculus — garantisce l'espressione della quantità causale come funzionale di $P(V)$. Ma non possono sostituire i dati interventistici per l'identificazione della struttura stessa: questa sostituzione è esattamente ciò che i teoremi di strettezza della gerarchia escludono.

Il secondo corollario riguarda il *tipo di architettura*. I sistemi che incorporano struttura causale in modo esplicito — modelli strutturali causali con grafo specificato da esperti del dominio o scoperto tramite algoritmi causali su dati sufficientemente ricchi di variazione interventistica, world models con azioni esplicite come variabili interventistiche, sistemi neurali con l'invarianza causale come vincolo di addestramento anziché come semplice regolarizzatore

statistico — hanno capacità causali strutturalmente distinte da quelle dei sistemi puramente associativi, indipendentemente dalla dimensione parametrica degli uni e degli altri. L'architettura non è epistemicamente neutrale rispetto alla gerarchia causale: sistemi con architetture diverse operano su livelli distinti della gerarchia, e il salto tra livelli richiede un cambiamento architetturale, non un incremento di scala. L'aggiunta di parametri a un sistema di livello-1 non produce un sistema di livello-2 per le stesse ragioni per cui l'aggiunta di campioni osservazionali non produce dati interventistici: la differenza è qualitativa, non quantitativa.

Questo secondo corollario ha conseguenze precise per il dibattito sulla scaling hypothesis — l'ipotesi che l'aumento della scala dei modelli linguistici (parametri, dati di addestramento, potenza computazionale) produca sistematicamente capacità qualitativamente nuove, tra cui la capacità di ragionamento causale. Il confine identificato dalla teoria causale è strutturalmente impenetrabile dalla sola scala: un sistema addestrato sulla distribuzione osservazionale del linguaggio, per quanto massiccio, apprende le regolarità correlazionali del corpus e non la struttura causale del mondo che il corpus descrive. Le prestazioni apparentemente causali di tali sistemi — la correttezza nel rispondere a quesiti causali semplici, la capacità di generare spiegazioni causalmente plausibili — sono il prodotto della correlazione tra il tipo di domanda causale e la struttura delle risposte corrette nel corpus di addestramento, non dell'applicazione di un meccanismo causale. La differenza si manifesta nei test di invarianza: quando la distribuzione di test differisce dalla distribuzione di addestramento per via di un intervento causale — una perturbazione del meccanismo generatore dei dati — un sistema associativo fallisce sistematicamente, mentre un sistema causalmente strutturato mantiene la sua accuratezza nella misura in cui il meccanismo perturbato non è quello su cui è condizionata la sua stima.

Il terzo corollario riguarda la *valutazione*. I benchmark correnti di valutazione della capacità causale dei sistemi di IA — comprese molte varianti dei dataset causali QA, i test di coerenza controfattuale, i dataset di commonsense causale — presentano un difetto strutturale: misurano la correttezza delle risposte a domande causali senza distinguere le risposte causalmente corrette da quelle correlazionalmente corrette nel corpus. Un sistema che risponde correttamente a "Se avessi preso l'ombrello, non mi sarei bagnato" sulla base del pattern statistico testo → risposta_attesa nella distribuzione di addestramento esibisce la medesima prestazione, sul benchmark standard, di un sistema che risponde sulla base di un genuino modello causale del meccanismo pioggia → bagnatura → ombrello. I due sistemi si distinguono sul benchmark di invarianza interventistica: se si introduce un nuovo contesto in cui il pattern statistico standard è assente ma la struttura causale è preservata, solo il secondo sistema mantiene le prestazioni.

Una valutazione causalmente informativa richiede pertanto: test di invarianza distributiva sotto shift causati da interventi espliciti, non da variazioni statistiche; test di composizione causale su combinazioni di meccanismi non visti durante l'addestramento, che verifichino la generalizzazione strutturale anziché la generalizzazione distribuzionale; test di contraffattuali con esiti contrastivi — situazioni in cui l'esito atteso secondo la distribuzione osservazionale contraddice l'esito causalmente corretto in virtù di un intervento. La costruzione di tali benchmark richiede la specifica esplicita della struttura causale di riferimento — il DAG del dominio considerato — e la generazione deliberata di casi che mettano alla prova l'identificabilità delle componenti causali rispetto alle associative. Questo è un programma di

valutazione sostanzialmente più costoso e teoricamente più esigente di quello attuale, ma è l'unico che misura ciò che si pretende di misurare.

Il quarto corollario riguarda le *applicazioni ad alto impatto*. Nelle applicazioni in cui le decisioni hanno conseguenze rilevanti e irreversibili — medicina di precisione, politiche di salute pubblica, decisioni giuridiche, politiche economiche, sistemi di credito e assicurazione — l'uso di sistemi di livello-1 per rispondere a domande di livello-2 o livello-3 non è soltanto epistemicamente scorretto ma produce errori sistematici con direzione determinata dalla struttura causale del dominio. Un sistema raccomanda trattamenti medici sulla base di correlazioni osservazionali invece di effetti causali — e il bias è del segno di $\Delta_{\text{confounding}}$, che dipende dalla struttura delle cause comuni tra trattamento ed esito. Suggerisce politiche pubbliche sulla base di associazioni storiche invece di meccanismi causali — e fallisce precisamente quando la politica interviene sui meccanismi, modificando le correlazioni su cui il sistema si fondava, come la legge di Goodhart predice. Valuta la probabilità controfattuale di un evento — "Se il paziente avesse ricevuto il trattamento alternativo, sarebbe guarito?" — senza avere accesso alla struttura funzionale necessaria per calcolare $P(Y_{\{x'\}} = y \mid X = x, Y = y')$ in modo corretto.

Questi errori non sono casuali ma sistematici, e la loro sistematicità implica che il loro impatto atteso può essere stimato — almeno in ordine di grandezza — dalla struttura causale del dominio. Per un confondimento latente di forza γ nella parametrizzazione della Parte Seconda, il bias atteso è

$$\mathbb{E}[\beta_{\text{OLS}} - \beta_{\text{true}}] = \gamma \cdot \sigma_U / \sigma_X \cdot \rho_{\{UX\}}$$

dove $\rho_{\{UX\}}$ è la correlazione tra il confounder latente e il trattamento. In applicazioni mediche con confonditori forti — indicazione per trattamento correlata con la prognosi — questo bias può facilmente essere dell'ordine di grandezza dell'effetto causale stesso, producendo decisioni sistematicamente errate. La consapevolezza di questa sistematicità non risolve il problema — per risolverlo occorrono dati interventistici o assunzioni identificative verificabili — ma impone la prudenza epistemica che è il minimo etico richiesto dal deploy in contesti ad alto impatto.

Il quinto corollario — spesso trascurato nelle discussioni pratiche — riguarda la *comunicazione dell'incertezza causale*. I sistemi di IA che operano in contesti decisionali non trasmettono soltanto previsioni ma anche, implicitamente, livelli di confidenza nelle previsioni. Un sistema che risponde a una domanda causale con un'unica stima puntuale, senza segnalare che quella stima è identificabile soltanto sotto assunzioni non verificabili, trasmette una falsa certezza che può essere epistemicamente più dannosa dell'ignoranza dichiarata. La teoria della trasportabilità — con la sua caratterizzazione formale delle condizioni in cui le conclusioni causali possono essere trasferite tra domini — fornisce un linguaggio preciso per comunicare i limiti: "questa stima è valida nel dominio di sorgente e trasportabile al dominio di destinazione se e soltanto se i meccanismi corrispondenti alle variabili di selezione $S_1, \dots, S_n$ sono condivisi tra i due domini". Questa precisione è più informativa, e quindi più utile al decisore, di un valore numerico accompagnato da un intervallo di confidenza statistico che non cattura l'incertezza strutturale.

V. Apertura: la complementarità tra statistica, causalità e conoscenza strutturale

Il titolo di questa monografia parla di *limiti strutturali* del ragionamento statistico. È necessario precisare, in sede di chiusura, che il termine "limite" non ha una valenza esclusivamente restrittiva: delimitare il territorio accessibile a uno strumento è, al tempo stesso, definire il territorio entro cui quello strumento esercita la sua sovranità. La statistica e l'analisi delle distribuzioni osservazionali sono strumenti di straordinaria potenza entro il loro dominio di validità — il livello-1 della gerarchia causale, la stima dei parametri causali una volta nota la struttura del DAG, l'inferenza bayesiana con priors strutturali motivati dalla teoria del dominio, la stima non parametrica delle distribuzioni post-intervento nei casi in cui il do-calculus garantisce l'espressione come funzionale di $P(V)$. I limiti strutturali non diminuiscono lo statuto epistemico della statistica: la collocano nella sua posizione corretta all'interno di un quadro più ampio in cui la struttura causale e la distribuzione osservazionale svolgono ruoli distinti e complementari.

Quel quadro più ampio è precisamente l'architettura causale sviluppata in questa monografia: una struttura a tre livelli in cui la statistica occupa il piano basale dell'associazione, l'inferenza interventistica il piano mediano dell'intervento, e il ragionamento controfattuale il piano superiore della retrodizione. I tre piani non sono in competizione ma in stretta interconnessione: l'identificazione causale usa le probabilità osservazionali per stimare i parametri dei modelli la cui struttura è fornita dalla teoria e dall'esperimento; l'inferenza interventistica usa quella struttura per correggere le stime statistiche del confondimento e calcolare $P(V \mid \text{do}(X = x))$; l'inferenza controfattuale usa entrambe per rispondere alle domande più profonde — sulla responsabilità, sulla controstoria, sulla personalizzazione delle decisioni — che la pratica clinica, giuridica e di policy continuamente avanza. Le tre modalità si alimentano reciprocamente: migliori stime statistiche producono parametri causali più accurati; una struttura causale meglio specificata produce stime statistiche corrette anziché distorte dal confondimento; un ragionamento controfattuale disciplinato produce valutazioni di policy che la sola statistica non può fornire.

Il programma dell'intelligenza artificiale causale — la costruzione di sistemi che operino genuinamente a livello-2 e livello-3 della gerarchia — non è pertanto soltanto un programma tecnico ma un programma epistemologico: richiede la reintegrazione della conoscenza qualitativa della struttura come componente esplicita e irrinunciabile dei sistemi computazionali, in complementarità con la capacità quantitativa di stimare distribuzioni, ottimizzare parametri e approssimare funzioni complesse. I teoremi di questa monografia mostrano che questa reintegrazione non è opzionale per chi intenda costruire sistemi con genuina capacità causale: è necessaria nel senso logico del termine — imposta dalla struttura della conoscenza causale stessa, non da preferenze metodologiche o filosofiche contingenti.

Questa necessità non è un ostacolo ma un orientamento. Sapere con precisione formale cosa non è possibile — non-identificabilità dal solo $P(V)$, impossibilità del disentanglement non supervisionato, non-trasportabilità in assenza di assunzioni sull'S-DAG — è epistemicamente più prezioso di un'indistinta fiducia nella potenza dei dati. Ogni teorema di impossibilità, nella tradizione della matematica e della logica, ha aperto nuovi programmi di ricerca indicando i confini entro cui la costruzione è possibile: il teorema di Gödel ha aperto la teoria della dimostrazione e la logica della prova; il teorema di Turing ha aperto la teoria della

complessità computazionale; la caratterizzazione di Shannon ha aperto la teoria dell'informazione. Il sistema di teoremi di impossibilità della teoria causale apre il programma dell'intelligenza artificiale strutturalmente causale: sistemi che scelgono esplicitamente le assunzioni causali su cui la loro inferenza si fonda, comunicano i limiti della loro validità, segnalano quando le domande poste eccedono la loro struttura identificativa, e si avvalgono di dati interventistici per espandere la frontiera del possibile nel dominio specifico di applicazione.

I contributi teorici che costituiscono le appendici di questa monografia — le derivazioni del do-calculus, la geometria informazionale della divergenza KL nei modelli causali, la riduzione da 3-colorazione per la NP-difficoltà dell'apprendimento strutturale, le condizioni di rango nei modelli lineari gaussiani, la formula generale di trasporto, le correzioni di Bekker per il regime degli strumenti deboli — non sono ornamenti tecnici ma strumenti pratici del programma. Le derivazioni del do-calculus mostrano con precisione quali ipotesi topologiche — d-separazioni specifiche nei grafi mutilati $G_{\setminus \{X\}}$, $G_{\setminus \{X, Z\}}$, $G_{\setminus \{X, Z(W(Z))\}}$ — sostengono ciascuna delle tre regole, consentendo di verificare caso per caso quando le regole si applicano. La geometria KL mostra come quantificare il costo dell'ignoranza strutturale in unità di informazione — una metrica concreta per valutare il valore atteso dell'acquisizione di conoscenza causale aggiuntiva in un problema applicativo. La teoria della trasportabilità e le correzioni di Bekker mostrano come la conoscenza strutturale si traduce in pratica statistica migliorata: la trasportabilità per il transfer learning causale, le correzioni di Bekker per l'inferenza robusta con strumenti deboli.

L'unità della monografia — dall'anomalia del paradosso di Simpson ai teoremi di impossibilità del causal representation learning, attraverso il do-calculus, la topologia delle classi di equivalenza, gli algoritmi di scoperta causale e la diagnosi dei sistemi di IA — non è soltanto narrativa ma matematica: ogni risultato è un nodo di una rete di implicazioni formali che ha come centro il teorema di strettezza della gerarchia causale e come conseguenza pratica il principio di responsabilità epistemica enunciato alla fine della Parte Sesta. Il confine tra ciò che un sistema sa e ciò che sembra di sapere è il confine tra la struttura causale che ha effettivamente appreso e le correlazioni statistiche che esibisce. Rendere quel confine visibile, misurabile e formalmente preciso è il contributo che questa monografia offre alla teoria e alla pratica dell'intelligenza artificiale causalmente fondata.

Appendice A: Il Do-calculus — Derivazioni Complete delle Tre Regole

A.1 Notazione e premesse

Le tre regole del do-calculus operano su distribuzioni di probabilità indicizzate da modelli causali strutturali (SCM) compatibili con un DAG G . La notazione è quella standard della Parte Seconda: $G_{\setminus \{X\}}$ indica il grafo G con gli archi entranti in X rimossi; $G_{\setminus \{X\}}$ indica G con gli archi uscenti da X rimossi; $(A \perp_G B \mid C)$ indica che A e B sono d-separati da C in

G. La distribuzione $P(Y \mid \text{do}(X=x), Z=z, W=w)$ è la distribuzione di Y nel modello mutilato $\text{do}(X=x)$ condizionata alle osservazioni $Z=z, W=w$.

Lemma fondamentale (indipendenza nel modello mutilato). Per ogni SCM M compatibile con G , se $(A \perp \{G\{X\}\} B \mid C)$ allora $A \perp B \mid C, \text{do}(X=x)$ in ogni modello mutilato M_x . Questo segue dal teorema di Markov causale applicato al modello mutilato: M_x è un SCM con grafo $G_{\{X\}}$, e la condizione di Markov implica che le d-separazioni in $G_{\{X\}}$ corrispondono a indipendenze condizionali in $P_{\{M_x\}}$.

A.2 Dimostrazione della Regola 1

Regola 1. Se $(Y \perp \{G\{X\}\} Z \mid X, W)$, allora:

$$P(Y \mid \text{do}(X=x), Z=z, W=w) = P(Y \mid \text{do}(X=x), W=w)$$

Dimostrazione. Per il lemma fondamentale, la condizione $(Y \perp \{G\{X\}\} Z \mid X, W)$ implica $Y \perp Z \mid X, W$ nel modello mutilato M_x . Dunque la distribuzione condizionale di Y dato $(\text{do}(X=x), Z=z, W=w)$ nel modello mutilato è:

$$P_{\{M_x\}}(Y \mid Z=z, W=w) = P_{\{M_x\}}(Y \mid W=w)$$

poiché la condizionamento su Z è irrilevante per Y quando X è fissato (tramite intervento) e W è osservata. Tradotto in notazione $\text{do}(\cdot)$:

$$P(Y \mid \text{do}(X=x), Z=z, W=w) = P(Y \mid \text{do}(X=x), W=w) \quad \square$$

Interpretazione. La Regola 1 è la *regola di cancellazione delle osservazioni irrilevanti*: se Z è d-separato da Y nel grafo mutilato $G_{\{X\}}$ dato (X, W) , l'aggiunta dell'osservazione $Z=z$ non modifica la distribuzione di Y dopo l'intervento $\text{do}(X=x)$.

A.3 Dimostrazione della Regola 2

Regola 2. Se $(Y \perp \{G\{X, Z\}\} Z \mid X, W)$, allora:

$$P(Y \mid \text{do}(X=x), \text{do}(Z=z), W=w) = P(Y \mid \text{do}(X=x), Z=z, W=w)$$

dove $G_{\{X, Z\}}$ è il grafo G con gli archi entranti in X e in Z rimossi.

Dimostrazione. La condizione $(Y \perp \{G\{X, Z\}\} Z \mid X, W)$ dice che Y è d-separato da Z in $G_{\{X, Z\}}$ dato (X, W) . Consideriamo il modello doppiamente mutilato $M_{\{x,z\}} = M$ con equazioni strutturali di X e Z sostituite dalle costanti x e z rispettivamente. Il grafo di $M_{\{x,z\}}$ è $G_{\{X, Z\}}$.

Per il teorema di Markov applicato a $M_{\{x,z\}}$ con grafo $G_{\{X, Z\}}$: la condizione di d-separazione $(Y \perp \{G\{X, Z\}\} Z \mid X, W)$ implica $Y \perp Z \mid X, W$ in $P_{\{M_{\{x,z\}}\}}$.

Ma in $M_{\{x,z\}}$, Z è fissata a z per intervento: $P_{\{M_{\{x,z\}}\}}(Z) = \delta_{\{Z=z\}}$. La condizionalizzazione su $Z=z$ nel modello $M_{\{x,z\}}$ è quindi equivalente a non condizionalizzare (Z è una costante). Dunque:

$$P(Y \mid \text{do}(X=x), \text{do}(Z=z), W=w) = P_{\{M_{\{x,z\}}\}}(Y \mid W=w)$$

D'altro canto, nel modello M_x (solo X mutilato), l'osservazione $Z=z$ introduce informazione causale attraverso il percorso $Z \rightarrow Y$ (se Z è genitore di Y). Ma la condizione $(Y \perp \{G\{X, Z\} \mid Z \mid X, W)$ garantisce che, nel grafo con anche gli archi entranti in Z rimossi, Y e Z sono d-separati. Questo implica che la distribuzione di Y dato $(\text{do}(X=x), Z=z, W=w)$ nel modello M_x è la stessa che nel modello $M_{\{x,z\}}$:

$$P(Y \mid \text{do}(X=x), Z=z, W=w) = P_{\{M_{\{x,z\}}\}}(Y \mid W=w) = P(Y \mid \text{do}(X=x), \text{do}(Z=z), W=w)$$

□

Interpretazione. La Regola 2 è la *regola di scambio osservazione-intervento*: quando la d-separazione nella struttura doppiamente mutilata è soddisfatta, osservare $Z=z$ equivale a intervenire su Z con $\text{do}(Z=z)$.

A.4 Dimostrazione della Regola 3

Regola 3. Sia $W(Z)$ l'insieme dei nodi in W che non sono antenati di alcun nodo in Z nel grafo $G_{\{X\}}$ (il grafo con archi entranti in X rimossi). Se $(Y \perp \{G\{X, Z(W(Z))\} \mid Z \mid X, W)$, dove $G_{\{X, Z(W(Z))\}}$ indica il grafo $G_{\{X\}}$ con gli archi uscenti da Z verso $W(Z)$ rimossi, allora:

$$P(Y \mid \text{do}(X=x), \text{do}(Z=z), W=w) = P(Y \mid \text{do}(X=x), W=w)$$

Dimostrazione (schema). La dimostrazione richiede di mostrare che nel modello mutilato M_x , l'intervento aggiuntivo $\text{do}(Z=z)$ non modifica la distribuzione di Y condizionata su $W=w$.

La condizione strutturale specifica i nodi $W(Z)$ — i nodi in W che non sono antenati di Z nel grafo $G_{\{X\}}$. Questi nodi non sono influenzati dall'intervento $\text{do}(Z=z)$: se $W_j \in W(Z)$, allora W_j non è discendente di Z in $G_{\{X\}}$, dunque la distribuzione di W_j non cambia quando si interviene su Z . Per i nodi $W_j \in W \setminus W(Z)$ — che sono antenati di Z nel grafo $G_{\{X\}}$ — il condizionamento su $W_j = w_j$ può aprire percorsi da Z verso Y attraverso la struttura di collider. La condizione $(Y \perp \{G\{X, Z(W(Z))\} \mid Z \mid X, W)$ cattura precisamente la richiesta che, anche dopo avere rimosso gli archi uscenti da Z verso $W(Z)$, Y rimanga d-separato da Z in $G_{\{X\}}$ dato (X, W) . Questo è sufficiente per garantire che $P_{\{M_x\}}(Y \mid Z=z, W=w) = P_{\{M_x\}}(Y \mid W=w)$, e dunque:

$$P(Y \mid \text{do}(X=x), \text{do}(Z=z), W=w) = P(Y \mid \text{do}(X=x), W=w) \quad \square$$

Interpretazione. La Regola 3 è la *regola di cancellazione degli interventi irrilevanti*: quando le condizioni topologiche sono soddisfatte, intervenire su Z non modifica la distribuzione di Y dopo $\text{do}(X=x)$.

Appendice B: Divergenza KL e Geometria dell'Informazione nei Modelli Causali

B.1 La divergenza di Kullback-Leibler

La divergenza di Kullback-Leibler (KL) tra due distribuzioni di probabilità P e Q su uno spazio misurabile $(\Omega, \mathcal{F})$ è definita come:

$$D_{\text{KL}}(P \parallel Q) = \int_{\Omega} \log(dP/dQ) \cdot dP = \mathbb{E}_P[\log(P(X)/Q(X))]$$

quando P è assolutamente continua rispetto a Q ($P \ll Q$); altrimenti $D_{\text{KL}}(P \parallel Q) = +\infty$. La divergenza KL non è una distanza nel senso metrico — non è simmetrica ($D_{\text{KL}}(P \parallel Q) \neq D_{\text{KL}}(Q \parallel P)$ in generale) e non soddisfa la disuguaglianza triangolare — ma è una misura fondamentale della discrepanza tra distribuzioni con molteplici interpretazioni: come eccesso di entropia, come test di ipotesi nulla (per il lemma di Stein), come guadagno di informazione bayesiana.

Proprietà fondamentali.

Non-negatività: $D_{\text{KL}}(P \parallel Q) \geq 0$, con uguaglianza se e solo se $P = Q$ quasi ovunque. Questa proprietà segue dalla disuguaglianza di Jensen applicata alla funzione convessa $-\log$:

$$D_{\text{KL}}(P \parallel Q) = -\mathbb{E}_P[\log(Q(X)/P(X))] \geq -\log(\mathbb{E}_P[Q(X)/P(X)]) = -\log(1) = 0$$

Additività per distribuzioni prodotto: Se $P = \otimes_i P_i$ e $Q = \otimes_i Q_i$ sono distribuzioni prodotto, allora $D_{\text{KL}}(P \parallel Q) = \sum_i D_{\text{KL}}(P_i \parallel Q_i)$.

Relazione con l'entropia: $D_{\text{KL}}(P \parallel Q) = H(P, Q) - H(P)$ dove $H(P, Q) = -\mathbb{E}_P[\log Q(X)]$ è l'entropia crociata e $H(P) = -\mathbb{E}_P[\log P(X)]$ è l'entropia di Shannon.

B.2 La divergenza KL nei modelli causali: quantificare il costo dell'ignoranza strutturale

Nel contesto dei modelli causali strutturali, la divergenza KL quantifica il costo dell'uso di un modello sbagliato per rispondere a domande causali. Siano M_1 e M_2 due SCM compatibili con lo stesso grafo G ma con parametri diversi, o compatibili con grafi diversi G_1 e G_2 . La divergenza KL tra le distribuzioni osservazionali $P_{\{M_1\}}(V)$ e $P_{\{M_2\}}(V)$ è:

$$D_{\text{KL}}(P_{\{M_1\}} \parallel P_{\{M_2\}}) = \mathbb{E}_{P_{\{M_1\}}}[\log(P_{\{M_1\}}(V)/P_{\{M_2\}}(V))]$$

Se $D_{\text{KL}}(P_{\{M_1\}} \parallel P_{\{M_2\}}) = 0$ i due modelli sono osservazionalmente identici — Markov-equivalenti nel senso della Parte Terza. La divergenza KL tra le distribuzioni *interventistiche* $P_{\{M_1\}}(V \mid \text{do}(X=x))$ e $P_{\{M_2\}}(V \mid \text{do}(X=x))$ può essere non nulla anche quando $D_{\text{KL}}(P_{\{M_1\}}(V) \parallel P_{\{M_2\}}(V)) = 0$: i modelli Markov-equivalenti possono avere distribuzioni interventistiche diverse. Questo è precisamente la misura quantitativa dell'ambiguità causale residua nella classe di equivalenza di Markov.

B.3 Informazione mutua e decomposizione in termini causali

L'informazione mutua tra X e Y è definita come:

$$I(X; Y) = D_{\text{KL}}(P(X, Y) \parallel P(X) \otimes P(Y)) = \mathbb{E}_{\{P(X, Y)\}} [\log(P(X, Y) / (P(X) \cdot P(Y)))]$$

Questa quantità misura la dipendenza statistica tra X e Y — esattamente una misura di livello-1. La decomposizione causale dell'informazione mutua distingue tra la parte che riflette una relazione causale diretta e la parte che riflette il confondimento:

$$I(X; Y) = I_{\text{causal}}(X \rightarrow Y) + I_{\text{confounding}}(X, Y \mid U)$$

dove $I_{\text{causal}}(X \rightarrow Y)$ è la riduzione di incertezza su Y dovuta alla conoscenza di X attraverso il percorso causale diretto $X \rightarrow Y$, e $I_{\text{confounding}}(X, Y \mid U)$ è la dipendenza che emerge dalla causa comune latente $X \leftarrow U \rightarrow Y$. Nel modello lineare gaussiano con i parametri della Parte Seconda:

$$I(X; Y) = -(1/2) \cdot \log(1 - \varrho_{XY}^2)$$

dove $\varrho_{XY} = [\beta(\alpha^2 + \sigma^2 X) + \alpha\gamma] / [\sqrt{(\alpha^2 + \sigma^2 X)} \cdot \sqrt{(\beta^2(\alpha^2 + \sigma^2 X) + (\beta\alpha + \gamma)^2 + \sigma^2 Y)}]$ è la correlazione di Pearson. La parte causale è:

$$I_{\text{causal}}(X \rightarrow Y) = -(1/2) \cdot \log(1 - [\beta \cdot \sigma_X / \sigma_Y]^2)$$

e la parte di confondimento è la differenza. Nei modelli senza confondimento ($\gamma = 0$): $I(X; Y) = I_{\text{causal}}(X \rightarrow Y)$ — tutta l'informazione mutua è causale. Nei modelli con confondimento: la differenza $I(X; Y) - I_{\text{causal}}(X \rightarrow Y)$ quantifica il "rumore informativo" causato dalla dipendenza spuria.

B.4 Il criterio di divergenza KL per la selezione del modello causale

Nel contesto della scoperta causale score-based (Appendice C), il punteggio BIC è un'approssimazione della log-verosimiglianza marginale:

$$\text{BIC}(G, D) \approx \log P(D \mid G) - (|G|/2) \cdot \log m \approx -m \cdot D_{\text{KL}}(P_D \parallel P_G) - (|G|/2) \cdot \log m$$

dove P_G è la distribuzione osservazionale del miglior modello in classe G e P_D è la distribuzione empirica dei dati. La minimizzazione di BIC corrisponde alla massimizzazione della verosimiglianza penalizzata per complessità, il che equivale a trovare il DAG G che minimizza la divergenza KL dall'empirica P_D tra tutti i DAG, penalizzando i grafi troppo complessi.

Appendice C: NP-difficoltà dell'Apprendimento Strutturale — Dimostrazione Dettagliata

C.1 La riduzione da 3-colorazione di grafi

Alternativamente alla riduzione da MAX-2SAT abbozzata nella Parte Quarta, presentiamo una riduzione dal problema della 3-colorazione di grafi, che ha una struttura più trasparente per i fini di questa Appendice.

Definizione (3-colorazione). Dato un grafo non orientato $H = (V_H, E_H)$, una 3-colorazione è una funzione $c : V_H \rightarrow \{1, 2, 3\}$ tale che per ogni arco $(u, v) \in E_H$: $c(u) \neq c(v)$. Il problema decisionale 3-COL chiede se H ammette una 3-colorazione; è NP-completo per riduzione da 3-SAT.

Costruzione della riduzione. Dato $H = (V_H, E_H)$ con $|V_H| = n$, costruiamo un dataset D e un punteggio $S(G, D)$ tali che il DAG che massimizza $S(G, D)$ esiste se e soltanto se H è 3-colorabile.

Per ogni nodo $v_i \in V_H$, introduciamo nel dataset una variabile categorica $V_i \in \{1, 2, 3\}$ e una variabile binaria ausiliaria W_i . Per ogni arco $(v_i, v) \in E_H$, introduciamo una variabile binaria E_i . Il dataset D è costruito come segue:

per ogni assegnazione di colori $c : V_H \rightarrow \{1, 2, 3\}$ valida (3-colorazione), si include nel dataset $m/|Col(H)|$ osservazioni con $V_i = c(v_i)$ per tutti gli i , $W_i = 1$ se $c(v_i) = 1$ e 0 altrimenti, e $E_i = 1$ se $c(v_i) \neq c(v)$ e 0 altrimenti.

Se H non è 3-colorabile, il dataset D è vuoto (nessuna 3-colorazione esiste) e nessun DAG produce il punteggio massimo.

Il punteggio $S(G, D)$ assegna punteggio alto ai DAG in cui le relazioni di dipendenza tra le variabili rispecchiano le strutture di 3-colorazione: dipendenze tra V_i e E_i che codificano il vincolo $c(v_i) \neq c(v)$. La costruzione esatta è tecnica e rimandiamo a Chickering (1996) per i dettagli.

C.2 Algoritmi approssimati e garanzie di approssimazione

La NP-difficoltà dell'apprendimento strutturale esatto motiva lo sviluppo di algoritmi approssimati. Per una classe ristretta di problemi — DAG con bounded treewidth — esistono algoritmi esatti con complessità pseudo-polinomiale; per la classe generale, le garanzie di approssimazione sono deboli.

Algoritmo MMHC (Max-Min Hill-Climbing). L'algoritmo MMHC di Tsamardinos et al. (2006) combina un fase di stima locale del Markov blanket (per inizializzare il grafo) con una fase di ricerca hill-climbing guidata dal punteggio BIC. La complessità è $O(n^2 \cdot m \cdot d)$ dove d

è il grado massimo stimato del grafo. Non ha garanzie di approssimazione rispetto al massimo globale del BIC, ma empiricamente produce soluzioni di buona qualità su dataset reali.

Algoritmi esatti con DP. Singh e Moore (2005) e Silander e Myllymäki (2006) hanno sviluppato algoritmi di programmazione dinamica che trovano il DAG ottimale per il punteggio BIC in tempo $O(2^n \cdot n)$. Per $n \leq 20-25$, questi algoritmi sono praticamente eseguibili; per $n > 30$ diventano impraticabili. La struttura della DP sfrutta la decomposizione del punteggio: $S(G, D) = \sum_i s(V_i, PA(V_i), D)$, e la programmazione dinamica trova l'ordinamento topologico ottimale esplorando tutti i 2^n sottoinsiemi di variabili.

Appendice D: Modelli Lineari Gaussiani — Identificabilità e Condizioni di Rango

D.1 Il sistema di equazioni strutturali lineari e la forma ridotta

In un modello causale strutturale lineare su n variabili endogene $V = (V_1, \dots, V_n)$ con equazioni:

$$V = B \cdot V + \Gamma \cdot U$$

dove $B \in \mathbb{R}^{n \times n}$ è la matrice triangolare inferiore dei coefficienti strutturali (con zeri sulla diagonale, per aciclicità) e $\Gamma \in \mathbb{R}^{n \times k}$ è la matrice dei coefficienti delle variabili esogene $U \sim N(0, I_k)$, la forma ridotta è:

$$V = (I - B)^{-1} \cdot \Gamma \cdot U =: A \cdot U$$

La matrice $A = (I - B)^{-1} \cdot \Gamma$ è la matrice della forma ridotta. La distribuzione di V è gaussiana multivariata:

$$V \sim N(0, \Sigma) \text{ con } \Sigma = A \cdot A^T = (I - B)^{-1} \cdot \Gamma \cdot \Gamma^T \cdot [(I - B)^{-1}]^T$$

Il problema di identificazione chiede: dato Σ (la varianza osservabile), è possibile recuperare B e Γ in modo unico? La risposta dipende dalla struttura di B e Γ e dalla presenza di variabili esogene distinte per diverse variabili endogene.

D.2 Condizione di identificabilità di ordine: la condizione di rango

Per il modello lineare simultaneo con k equazioni (Σ è la matrice di varianza-covarianza delle variabili esogene, non necessariamente l'identità), la condizione classica di identificabilità di ordine e rango di Koopmans e Reiersøl (1950) è:

Condizione di ordine: l'equazione strutturale i -esima contiene almeno $n - 1$ variabili escluse (variabili presenti in altre equazioni ma assenti nell'equazione i).

Condizione di rango: la sottomatrice della matrice B ottenuta selezionando le righe delle variabili incluse nell'equazione i e le colonne delle variabili escluse dall'equazione i ha rango $n - 1$.

La condizione di ordine è necessaria per l'identificabilità; la condizione di rango è necessaria e sufficiente (per il caso esattamente identificato). Per il caso sovra-identificato (più restrizioni del necessario), il modello è identificabile se e soltanto se la condizione di rango è soddisfatta per ogni equazione.

D.3 Il modello con errori correlati: gli archi bidirezionali come parametri

Quando alcune equazioni strutturali condividono variabili esogene latenti — corrispondenti agli archi bidirezionali nel grafo ADMG — la matrice Σ degli errori strutturali non è diagonale. Sia $\Psi = \Gamma \cdot \Gamma^T$ la matrice di varianza degli errori strutturali — non diagonale in presenza di cause latenti comuni — e sia $\Sigma = (I-B)^{-1} \cdot \Psi \cdot [(I-B)^{-1}]^T$ la varianza osservabile.

In questo caso, il problema di identificazione è risolvere $\Sigma = (I-B)^{-1} \cdot \Psi \cdot [(I-B)^{-1}]^T$ per (B, Ψ) dati Σ . Questa è un'equazione matriciale in $n(n-1)/2 + n(n+1)/2 = n^2$ incognite con $n(n+1)/2$ equazioni (gli elementi distinti di Σ). Il sistema è generalmente sottodeterminato, confermando la non-identificabilità discussa nella Parte Seconda.

Le condizioni di identificabilità per modelli lineari Gaussiani con errori correlati — modelli LISREL nella tradizione della psicometria strutturale di Jöreskog (1970) — richiedono vincoli aggiuntivi sulla struttura di Ψ : zeri in posizioni specifiche della matrice degli errori, corrispondenti all'assunzione che certe variabili non condividano variabili latenti. Questi vincoli devono essere motivati dalla teoria del dominio — non possono essere imposti arbitrariamente senza alterare le conclusioni causali.

D.4 Casi particolari identificabili: la condizione di Bekker

Un caso speciale con risultati di identificabilità più forti è il modello con variabili strumentali. Se esiste una variabile Z_i (strumento) tale che Z_i causa V (rilevanza) ma non causa direttamente V attraverso percorsi non passanti per V (esclusione), allora il coefficiente causale $V \rightarrow V$ è identificabile tramite la formula IV (variabile strumentale):

$$\beta_{jk} = \text{Cov}(Z_i, V) / \text{Cov}(Z_i, V)$$

La condizione di esclusione — $\text{Cov}(Z_i, \varepsilon) = 0$ — è cruciale: se lo strumento influenza V anche attraverso percorsi diretti non passanti per V , l'estimatore IV è distorto. La verifica della condizione di esclusione richiede la teoria del dominio — non è testabile statisticamente. Bekker (1994) ha derivato le condizioni di identificabilità per sistemi con strumenti multipli, mostrando che la sovra-identificazione (più strumenti del necessario) permette test di validità degli strumenti tramite il test di Sargan-Hansen, che verifica la consistenza delle stime IV da strumenti diversi.

Appendice E: Trasportabilità e la Formula di Trasporto

E.1 Il problema del trasferimento causale

La trasportabilità è il problema di trasferire conclusioni causali tra domini che differiscono per la distribuzione di alcune variabili o per alcune delle loro relazioni causali. Nella pratica scientifica e ingegneristica, ogni applicazione di un modello causale al di fuori del contesto in cui è stato stimato è un problema di trasportabilità: i trial clinici sono condotti su popolazioni selezionate e si vogliono generalizzare alla popolazione generale; gli studi economici in un paese si vogliono applicare a un altro; i sistemi di IA addestrati in un ambiente operativo vengono deployati in ambienti diversi.

La difficoltà del trasferimento causale è strutturale: anche se la struttura causale qualitativa è la stessa nei due domini (stesso grafo G), le distribuzioni dei confonditori, dei predittori e dei mediatori possono differire in modi che rendono le stime causali non direttamente trasferibili. La semplice "portabilità" — applicare la stessa formula stimata nella sorgente alla destinazione — produce stime distorte se le variabili di confondimento hanno distribuzioni diverse tra sorgente e destinazione.

E.2 Il formalismo degli S-DAG

La teoria della trasportabilità di Pearl e Bareinboim (2011, 2012) introduce le variabili di selezione S per rappresentare le differenze tra domini nel DAG aumentato.

Definizione (S-DAG). Un S-DAG $\langle G, S \rangle$ è un DAG G aumentato con variabili di selezione $S = (S_1, \dots, S_k)$ tali che: ogni S_i è un nodo radice (senza genitori) nel S-DAG; ogni S_i è genitore di almeno un nodo $V \in V$; $S_i \notin V$. L'arco $S_i \rightarrow V$ indica che il meccanismo causale di V — l'equazione strutturale f — differisce tra il dominio sorgente ($S_i = 0$) e il dominio di destinazione ($S_i = 1$).

La distribuzione nel dominio sorgente corrisponde alla distribuzione del modello condizionata su $S = 0$; la distribuzione nel dominio di destinazione corrisponde alla distribuzione condizionata su $S = 1$.

Esempi di differenze tra domini.

Perturbazioni della distribuzione delle esogene: se la distribuzione di una variabile X cambia tra sorgente e destinazione (per esempio, diversa distribuzione demografica), senza che il meccanismo causale di X cambi, l'arco di selezione è $S_i \rightarrow U_X$, dove U_X è la variabile esogena di X . Il meccanismo f_x rimane lo stesso; cambia solo $P(U_X)$.

Perturbazioni dei meccanismi causali: se l'effetto causale di X su Y è diverso tra i domini (per esempio, l'efficacia di un farmaco dipende da caratteristiche genetiche della popolazione che differiscono), l'arco di selezione è $S_i \rightarrow Y$, indicando che f_y differisce. Il meccanismo stesso cambia tra i domini.

Perturbazioni strutturali: in alcuni casi, il grafo stesso differisce tra i domini — esistono relazioni causali in un dominio che non esistono nell'altro. Questa è la forma più drastica di differenza tra domini e richiede assunzioni strutturali più forti per la trasportabilità.

E.3 Trasportabilità diretta e aggiustata

Definizione (trasportabilità diretta). La quantità causale $R = P_\Pi(Y|\text{do}(X=x))$ è direttamente trasportabile da Σ a Π se $R = P_\Sigma(Y|\text{do}(X=x))$ — i due domini hanno lo stesso valore per R . Una condizione sufficiente per la trasportabilità diretta è che nessuna variabile di selezione sia antenata di X o Y nel S-DAG.

Proposizione (condizione grafica per la trasportabilità diretta). Se nessuna variabile di selezione S_i ha percorso diretto verso X o Y nel S-DAG (rispettando la direzione degli archi), allora $P_\Sigma(Y|\text{do}(X=x)) = P_\Pi(Y|\text{do}(X=x))$.

La dimostrazione è immediata: se nessuna S_i è antenata di X o Y , allora il meccanismo causale f_y e le distribuzioni delle variabili nel percorso $\text{do}(X) \rightarrow Y$ sono le stesse in entrambi i domini. Le variabili di selezione non entrano nel calcolo dell'effetto interventistico.

Definizione (trasportabilità con aggiustamento). La quantità causale R è trasportabile con aggiustamento rispetto alle variabili $W \subseteq V$ se:

$$P_\Pi(Y|\text{do}(X=x)) = \sum_w P_\Sigma(Y|\text{do}(X=x), W=w) \cdot P_\Pi(W=w)$$

La formula dice: si usa la relazione causale quantitativa $P_\Sigma(Y|\text{do}(X=x), W=w)$ stimata nella sorgente, ma si pesa con la distribuzione $P_\Pi(W=w)$ della destinazione. Il set W cattura le differenze sistematiche tra i domini che distorcerebbero la trasportabilità diretta.

Identificazione delle variabili di aggiustamento. Le variabili di aggiustamento W devono soddisfare due condizioni: (1) W d-separa le variabili di selezione da Y nel S-DAG dopo il $\text{do}(X)$; (2) $P_\Pi(W)$ è nota o stimabile nella destinazione. La seconda condizione richiede che le variabili W siano osservabili nella destinazione — tipicamente soddisfatta per covariate demografiche, ma potenzialmente non soddisfatta per biomarker o variabili di contesto non misurate.

E.4 Il teorema di completezza della trasportabilità

Teorema (completezza della trasportabilità, Pearl e Bareinboim, 2011). Il do-calculus è completo per la trasportabilità: una quantità causale R è trasportabile dalla famiglia di sorgenti $\delta = \{P_\Sigma, P_\Sigma^{\text{do}(\cdot)}\}$ alla destinazione Π se e solo se R è calcolabile da δ usando le tre regole del do-calculus applicate all'S-DAG.

La dimostrazione di completezza usa la costruzione di un grafo hedge generalizzato che testimonia l'impossibilità: se l'algoritmo TR (Transportation) fallisce, il suo fallimento corrisponde all'esistenza di due SCM con lo stesso S-DAG e la stessa δ ma diversi valori di R nella destinazione — impossibilità costruttiva.

E.5 La formula generale di trasporto

In presenza di dati da K domini sorgente $\Sigma_1, \dots, \Sigma_K$ con S-DAG noti, la formula di trasporto generalizzata ha la forma:

$$P_\Pi(Y=y \mid \text{do}(X=x)) = \sum_v \prod \{V_i \in V \setminus \{X, Y\} \mid f_{\{i, \text{source}(i)\}}(v_i \mid \text{pa}(v_i))\} \cdot P_\Pi(U)$$

dove $\text{source}(i) \in \{\Sigma_1, \dots, \Sigma_K, \Pi\}$ indica quale dominio fornisce la stima per il meccanismo f_i , e $P_\Pi(U)$ è la distribuzione delle variabili esogene nella destinazione.

La formula chiarisce la struttura dell'aggregazione dei dati multi-sorgente: ogni meccanismo causale f_i è stimato nel dominio in cui è disponibile e osservabile, e i meccanismi sono poi assemblati nella destinazione usando la distribuzione target delle variabili esogene. La chiave è la modularità: se i meccanismi sono autonomi, si possono stimare separatamente e poi ricombinare.

E.6 Trasportabilità e generalizzazione dei trial clinici

La generalizzazione dei trial clinici randomizzati (RCT) dalla popolazione del trial alla popolazione target è il principale campo di applicazione della teoria della trasportabilità. Sia Σ la popolazione del trial (randomizzata, con dati sperimentali $P_\Sigma(Y \mid \text{do}(T))$) e Π la popolazione target (non randomizzata, con soli dati osservazionali $P_\Pi(V)$).

La formula di standardizzazione per la generalizzazione da RCT è:

$$P_\Pi(Y \mid \text{do}(T=t)) = \sum_w P_\Sigma(Y \mid \text{do}(T=t), W=w) \cdot P_\Pi(W=w)$$

dove W è il set di covariate di moderazione — variabili che modificano l'effetto del trattamento T su Y in modo che differisce tra la popolazione del trial e quella target.

La condizione di trasportabilità corrisponde all'assenza di interazione non catturata tra le variabili di selezione e l'effetto del trattamento: se l'efficacia del trattamento è la stessa in tutti i strati definiti da W nella sorgente, e la distribuzione di W nella destinazione è nota, la formula di standardizzazione è valida. Se l'efficacia differisce per ragioni non catturate da W — eterogeneità d'effetto non misurata — la standardizzazione non è corretta.

Il problema della generalizzazione degli RCT è un caso in cui la teoria causale formale ha impatti normativi: la pratica medica attuale spesso applica i risultati degli RCT senza aggiustamento per le differenze tra la popolazione del trial e quella target, assumendo implicitamente la trasportabilità diretta. Quando questa assunzione è violata — tipicamente per popolazioni anziane, donne in età fertile o bambini sottorappresentati nei trial — l'efficacia reale può differire sostanzialmente dall'efficacia stimata nel trial.

Appendice F: Strumenti Deboli e Correzioni di Bekker

F.1 Il problema degli strumenti deboli

Il metodo delle variabili strumentali (IV), introdotto nella Parte Prima come approssimazione pratica degli interventi nel framework causale, si basa sull'assunzione che lo strumento Z sia sufficientemente correlato con la variabile endogena X da permettere la stima dell'effetto causale di X su Y . Quando questa correlazione è debole, lo stimatore IV soffre di un bias di piccolo campione e di distribuzioni campionarie non standard che rendono i test di ipotesi standard non validi. Questo è il problema degli strumenti deboli, sistematizzato da Stock, Wright e Yogo (2002) e corretto nel framework di Bekker (1994).

Definizione formale della debolezza. Nel modello IV lineare bivariato:

$$Y = \beta X + u, \quad X = \gamma Z + v$$

lo strumento Z è *debole* se la correlazione parziale $\rho_{\{XZ|W\}}$ — misurata dopo aver controllato per le covariate aggiuntive W — è piccola nel senso preciso che il parametro di non centralità della statistica F per la significatività di Z nel primo stadio è piccolo rispetto alla dimensione campionaria:

$$\mu^2 = \gamma^2 \cdot \sigma_Z^2 / \sigma_v^2 \cdot m \rightarrow c$$

dove c è una costante finita invece di divergere all'infinito come richiesto per la convergenza in distribuzione standard.

Quando μ^2 è piccolo, lo stimatore IV $\beta_{\{IV\}} = (Z'X)^{-1} Z'Y$ non è asintoticamente normale con media β e varianza $\sigma^2/(m \cdot \gamma^2 \cdot \sigma_Z^2)$: la distribuzione asintotica è deformata, non centrata, e i test t e F standard sono invalidi.

F.2 La statistica F del primo stadio e la regola del pollice

Stock e Yogo (2005) derivano una regola pratica per valutare la robustezza degli stimatori IV alla debolezza degli strumenti: la statistica F del primo stadio — la statistica F per il test di

significatività congiunta degli strumenti nella regressione $X = \gamma Z + v$ — deve superare una soglia critica affinché il bias relativo dello stimatore IV sia inferiore a una soglia prestabilita.

Per un singolo strumento e un singolo endogeno, la soglia critica è:

$F > 10$ per garantire che il bias relativo $|\mathbb{E}[\beta_{IV}] - \beta| / \beta$ sia inferiore al 10% del bias OLS.

Questa "regola del 10" di Staiger e Stock (1997) è diventata standard pratico nella letteratura econometrica. Tuttavia, la regola è derivata per specifici valori dei parametri (modelli con omoschedasticità e un singolo strumento) e non vale in generale per modelli con molti strumenti, eteroschedasticità, o clustering. Le soglie corrette dipendono dal numero di strumenti e dalla struttura dell'eteroschedasticità.

Generalizzazione con più strumenti. Con K strumenti e L variabili endogene, Stock e Yogo (2005) derivano soglie critiche $\tau(K, L, \rho_{\max}, \alpha)$ per la statistica di Cragg-Donald:

$$CD = (m - K) / K \cdot \lambda_{\min}(\Xi)$$

dove $\lambda_{\min}$ è l'autovalore minimo di una matrice che misura la correlazione degli strumenti con le variabili endogene dopo concentrazione. La statistica di Cragg-Donald generalizza la statistica F del primo stadio al caso multivariato.

Le soglie critiche crescono sia in K sia in L : più strumenti richiedono correlazioni più forti per evitare il bias da debolezza. Per $K=1, L=1$: soglia ≈ 10 . Per $K=5, L=1$: soglia ≈ 22 . Per $K=10, L=2$: soglia ≈ 31 .

F.3 Inferenza robusta alla debolezza: il test di Anderson-Rubin

Lo stimatore IV β_{IV} è invalido sotto strumenti deboli. Ma è possibile costruire test di ipotesi su β che siano validi — con livello di tipo I corretto — indipendentemente dalla forza degli strumenti, anche se la potenza del test decresce con la debolezza.

Il test di Anderson-Rubin (AR, 1949) è il prototipo di tali test robusti. Per testare $H_0: \beta = \beta_0$, si calcola la statistica:

$$AR(\beta_0) = (Y - \beta_0 \cdot X)' P_Z (Y - \beta_0 \cdot X) / [(Y - \beta_0 \cdot X)' M_Z (Y - \beta_0 \cdot X) / (m - K)]$$

dove $P_Z = Z(Z'Z)^{-1}Z'$ e $M_Z = I - P_Z$ sono le matrici di proiezione ortogonale e supplementare sullo spazio di Z . Sotto H_0 e le assunzioni IV standard ($u \perp Z$), $AR(\beta_0) \sim F(K, m-K)$ — esattamente, non asintoticamente — indipendentemente dalla forza degli strumenti. Quindi il test AR mantiene il livello nominale α per qualsiasi γ , anche $\gamma = 0$.

La regione di confidenza di Anderson-Rubin al livello $1-\alpha$ è:

$$CR_{\{AR\}} = \{\beta_0 : AR(\beta_0) \leq F_{\{\alpha, K, m-K\}}\}$$

Questa regione è valida — ha probabilità di copertura esattamente $1-\alpha$ per qualsiasi valore di γ — ma può avere forma anomala: può essere vuota (se i dati contraddicono qualsiasi valore di β) o essere l'intera retta reale (se gli strumenti non hanno alcun potere discriminante su β).

F.4 Le correzioni di Bekker

Bekker (1994) affronta il problema degli strumenti deboli nel regime opposto: non strumenti singoli debolmente correlati con l'endogeno, ma molti strumenti moderatamente correlati. Il regime di molti strumenti deboli è rilevante in diverse applicazioni: l'uso di variabili lag multipli come strumenti nei panel dinamici, l'uso di caratteristiche geografiche granulari come strumenti in modelli di diffusione spaziale, l'uso di interazioni di variabili come strumenti nella letteratura di design of experiments.

L'asintosi di Bekker. Invece dell'asintotica standard dove K è fisso e $m \rightarrow \infty$ — che richiede $F \rightarrow \infty$ per la consistenza — Bekker considera l'asintotica dove sia m sia K crescono con il rapporto $K/m \rightarrow \alpha \in (0,1)$. In questo regime, lo stimatore IV standard (TSLS, Two-Stage Least Squares) soffre di bias proporzionale ad α :

$$\mathbb{E}[\beta_{\text{TSLS}}] \approx \beta + \alpha \cdot (\beta_{\text{OLS}} - \beta)$$

Il bias è una media pesata tra il valore vero β e il bias OLS β_{OLS} , con peso $\alpha = K/m$ che misura il grado di sovra-identificazione relativa alla dimensione del campione. Per K piccolo rispetto a m , $\alpha \approx 0$ e il bias è trascurabile — la situazione standard. Per $K = m/2$ (metà delle osservazioni è strumenti), $\alpha = 1/2$ e il bias è metà del bias OLS — sostanziale.

Lo stimatore LIML corretto di Bekker. Lo stimatore LIML (Limited Information Maximum Likelihood) ha un'asintotica di Bekker più favorevole:

$$\sqrt{m}(\beta_{\text{LIML}} - \beta) \rightarrow_d N(0, \sigma^2_{\text{LIML}})$$

con varianza σ^2_{LIML} dipendente da α , anche nel regime di Bekker. Il LIML non è consistente sotto l'asintotica di Bekker — ha un bias che cresce con α — ma la sua distribuzione asintotica è normale, rendendo i test di ipotesi standard asintoticamente validi nel senso che il livello effettivo converge al nominale.

Lo stimatore Fuller. Fuller (1977) propone una modificazione del LIML che riduce il bias attraverso una deflazione del denominatore:

$$\beta_{\text{Fuller}}(a) = \beta_{\text{LIML}} + a \cdot (\beta_{\text{OLS}} - \beta_{\text{LIML}}) / (m - K)$$

Per $a = 1$, il bias del Fuller è $O(m^{-1})$ rispetto al bias $O(K/m)$ del TSLS nel regime di Bekker. Il valore $a = 4$ produce lo stimatore con il minimo errore quadratico medio nel regime di molti strumenti secondo i calcoli di Bekker.

Le correzioni di Bekker per gli errori standard. Anche se β è stimato con bias ridotto, le formule standard per gli errori standard TSLS e LIML sono invalide nel regime di Bekker. Bekker (1994) deriva gli errori standard robusti:

$$\text{Var}_{\text{Bekker}}(\beta_{\text{LIML}}) = (1 - \alpha) \cdot \sigma^2_u / (m \cdot \gamma^2 \cdot \sigma^2_Z / \sigma^2_v)$$

che differiscono dagli errori standard standard LIML per il fattore $(1 - \alpha)$. Per α piccolo ($K \ll m$), la differenza è trascurabile. Per $\alpha = 0.5$, gli errori standard di Bekker sono il doppio degli errori standard standard — test basati sugli errori standard standard rifiuterebbero H_0 il doppio del necessario, producendo un eccesso di falsi positivi.

F.5 Connessione con il framework causale

Le correzioni di Bekker sono rilevanti non solo come tecnica econometrica ma come caratteristica strutturale del problema dell'identificazione causale tramite IV.

Nello schema causale della Parte Prima, lo strumento Z deve soddisfare tre condizioni: (1) $Z \rightarrow X$ (rilevanza: Z causa X); (2) $Z \perp U$ (esogeneità: Z è non correlato al confounder); (3) Z non ha effetti diretti su Y al di fuori del percorso $Z \rightarrow X \rightarrow Y$ (esclusione). Queste tre condizioni formano il nucleo identificante: permettono di estrarre la variazione di X causata da Z — una variazione "pulita" non contaminata dal confounder U .

La debolezza degli strumenti corrisponde alla condizione (1) debole: Z causa X , ma l'effetto causale è piccolo. Il bias IV è la conseguenza statistica di questa debolezza: quando la variazione di X causata da Z è piccola rispetto alla variazione totale di X , la stima IV diventa molto sensibile agli errori di misura e alla violazione dell'esogeneità.

Causalmente, questo significa che lo strumento Z *identifica* β solo per il sotto-gruppo della popolazione in cui Z ha effetto causale su X — l'interpretazione LATE (Local Average Treatment Effect) di Imbens e Angrist (1994). Per strumenti deboli, questo sotto-gruppo è piccolo e non necessariamente rappresentativo della popolazione generale. Il LATE stimato da strumenti deboli può essere molto diverso dall'ATE (Average Treatment Effect) sulla popolazione intera, e la stima è altamente variabile perché il sotto-gruppo rilevante è una piccola frazione del campione totale.

Le correzioni di Bekker sono quindi tecnicamente la risposta al problema statistico della debolezza, ma la risposta causalmente più informativa è quella di interpretare sempre la stima IV come LATE sull'insieme dei complier — il sotto-gruppo per cui Z ha effetto causale su X — e di valutare la rappresentatività di questo sotto-gruppo rispetto alla domanda causale di interesse.

Nota Bibliografica

La seguente nota bibliografica è organizzata per nuclei tematici, con indicazione del contributo specifico di ciascun lavoro al quadro teorico della monografia.

Fondamenti storici della causalità. La quadripartizione aristotelica delle cause è in *Fisica* II.3 (194b16–195b30) e *Metafisica* Δ.2 (1013a24–1014a25); la problematica humiana dell'inferenza causale è distribuita tra il *Trattato della natura umana* (1739), Libro I Parte III, e la *Ricerca sull'intelletto umano* (1748), Sezione IV–VII. I cinque canoni di Mill sono esposti in *Sistema di logica* (1843), Libro III Capitoli VIII–X. Wright (1921, *Correlation and causation*, *Journal of Agricultural Research* 20, 557–585) introduce il path analysis; Haavelmo (1943, *The*

statistical implications of a system of simultaneous equations, *Econometrica* 11, 1–12) formalizza il modello strutturale; Koopmans e Reiersøl (1950, *The identification of structural characteristics*, *Annals of Mathematical Statistics* 21, 165–181) sviluppano la teoria dell'identificabilità nei modelli simultanei lineari.

Il framework dei risultati potenziali: Neyman e Rubin. La notazione dei risultati potenziali è introdotta in Neyman (1923, *Sur les applications de la théorie des probabilités aux expériences agricoles: essai des principes*, *Roczniki Nauk Rolniczych* 10, 1–51; traduzione inglese in *Statistical Science* 5, 1990, 465–472) nel contesto degli esperimenti agrari randomizzati. L'estensione sistematica ai problemi di inferenza causale non sperimentale è opera di Rubin (1974, *Estimating causal effects of treatments in randomized and nonrandomized studies*, *Journal of Educational Psychology* 66, 688–701; 1978, *Bayesian inference for causal effects: the role of randomization*, *Annals of Statistics* 6, 34–58). L'assunzione SUTVA (Stable Unit Treatment Value Assumption) è formalizzata in Rubin (1980, *Randomization analysis of experimental data: the Fisher randomization test*, *Journal of the American Statistical Association* 75, 591–595). Holland (1986, *Statistics and causal inference*, *Journal of the American Statistical Association* 81, 945–960) fornisce una rassegna sistematica del framework e introduce la distinzione tra "effetti di cause" e "cause di effetti". L'equivalenza formale tra il framework dei potential outcomes e il formalismo di Pearl è dimostrata in Pearl (2000, capitolo 7) e in Heckman e Vytlačil (2005, *Structural equations, treatment effects and econometric policy evaluation*, *Econometrica* 73, 669–738). Imbens e Rubin (2015, *Causal Inference for Statistics, Social and Biomedical Sciences*, Cambridge University Press) è il riferimento standard per l'uso applicativo del framework potential outcomes in economia e medicina.

Metodi quasi-sperimentali e variabili strumentali. La formalizzazione dell'approccio a variabili strumentali nell'econometria classica è in Haavelmo (1943) e Koopmans e Reiersøl (1950), già citati. L'interpretazione LATE della stima IV per variabili binarie è introdotta in Imbens e Angrist (1994, *Identification and estimation of local average treatment effects*, *Econometrica* 62, 467–475), che dimostrano che la stima IV identifica l'ATE soltanto per il sottogruppo dei compliers quando il trattamento è eterogeneo. Angrist, Imbens e Rubin (1996, *Identification of causal effects using instrumental variables*, *Journal of the American Statistical Association* 91, 444–455) sistematizzano l'interpretazione LATE nel framework potential outcomes. La regressione con discontinuità (RDD, Regression Discontinuity Design) è introdotta da Thistlethwaite e Campbell (1960, *Regression-discontinuity analysis: an alternative to the ex post facto experiment*, *Journal of Educational Psychology* 51, 309–317) e sistematizzata da Hahn, Todd e Van der Klaauw (2001, *Identification and estimation of treatment effects with a regression-discontinuity design*, *Review of Economic Studies* 68, 201–209). Il metodo delle differenze-nelle-differenze ha i suoi fondamenti in Ashenfelter (1978, *Estimating the effect of training programs on earnings*, *Review of Economics and Statistics* 60, 47–57) e Card e Krueger (1994, *Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania*, *American Economic Review* 84, 772–793).

Strumenti deboli: teoria e correzioni. Il problema degli strumenti deboli è diagnosticato da Nelson e Startz (1990, *The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one*, *Journal of Business* 63, S125–S140). Staiger e Stock (1997, *Instrumental variables regression with weak instruments*, *Econometrica* 65, 557–586) introducono la statistica F del primo stadio come diagnostica standard e sviluppano l'asintotica con strumenti deboli (K fisso, concentrazione parametrica $\lambda = \alpha \cdot \sqrt{m}$ finita). Il test di Anderson-Rubin per l'ipotesi nulla sul coefficiente strutturale robusto alla debolezza è in Anderson e Rubin (1949,

Estimation of the parameters of a single equation in a complete system of stochastic equations, Annals of Mathematical Statistics 20, 46–63). Fuller (1977, *Some properties of a modification of the limited information estimator*, Econometrica 45, 939–953) introduce lo stimatore Fuller(a) come correzione del LIML con bias ridotto. Bekker (1994, *Alternative approximations to the distributions of instrumental variable estimators*, Econometrica 62, 657–681) sviluppa l'asintotica con molti strumenti ($K \rightarrow \infty$ con $K/m \rightarrow \alpha$) e le correzioni per gli errori standard. Donald e Newey (2001, *Choosing the number of instruments*, Econometrica 69, 1161–1191) estendono l'analisi di Bekker allo scenario di selezione del numero di strumenti. Mikusheva (2010, *Robust confidence sets in the presence of weak instruments*, Journal of Econometrics 157, 236–247) fornisce una rassegna dei metodi di inferenza robusti alla debolezza, comparando Anderson-Rubin, Moreira CLR (Conditional Likelihood Ratio) e altri.

Maximal Ancestral Graphs e grafi misti. Il formalismo degli Acyclic Directed Mixed Graphs (ADMG) e dei MAG è introdotto in Richardson (1996, *Discovering cyclic causal structure*, Technical Report CMU-CALD-96-057) per grafi potenzialmente ciclici e sistematizzato per il caso aciclico in Richardson e Spirtes (2002, *Ancestral graph Markov models*, Annals of Statistics 30, 962–1030) — il riferimento fondativo del formalismo. La teoria della m-separazione e la sua equivalenza con la d-separazione nella DAG originale è dimostrata in Richardson e Spirtes (2002). La caratterizzazione dell'equivalenza di Markov per MAG tramite scheletro, v-strutture e discriminating paths è in Zhang (2008, *Causal reasoning with ancestral graphs*, Journal of Machine Learning Research 9, 1437–1474), che dimostra anche la completezza delle dieci regole di orientazione FCI. L'algoritmo RFCI è in Colombo, Maathuis, Kalisch e Richardson (2012, *Learning high-dimensional directed acyclic graphs with latent and selection variables*, Annals of Statistics 40, 294–321). La completezza dell'algoritmo ID per ADMG — la generalizzazione del do-calculus ai modelli semi-markoviani — è in Shpitser e Pearl (2006, *Identification of joint interventional distributions in recursive semi-Markovian causal models*, AAAI 2006, 1219–1226). Il formalismo degli archi bidirezionali come rappresentazione del confondimento latente nei modelli strutturali lineari è studiato da Drton, Foygel e Sullivan (2011, *Global identifiability of linear structural equation models*, Annals of Statistics 39, 865–886) nel contesto dell'identificabilità algebrica.

Trasportabilità e transfer causale. La teoria formale della trasportabilità è introdotta in Pearl e Bareinboim (2011, *Transportability of causal and statistical relations: a formal approach*, AAAI 2011, 247–254) e sviluppata sistematicamente in Bareinboim e Pearl (2013, *A general algorithm for deciding transportability of experimental results*, Journal of Causal Inference 1, 107–134). Il formalismo degli S-DAG (Selection Diagrams, o DAG con variabili di selezione) e il teorema di completezza dell'algoritmo TR sono in Bareinboim e Pearl (2012, *Transportability from multiple environments with limited experiments*, NeurIPS 2012, 136–144). La generalizzazione a fonti multiple eterogenee è in Bareinboim e Pearl (2016, *Causal inference and the data-fusion problem*, Proceedings of the National Academy of Sciences 113, 7345–7352). L'applicazione della teoria alla generalizzazione dei trial clinici — il problema del trasferimento dalla popolazione dello studio alla popolazione bersaglio — è in Cole e Stuart (2010, *Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial*, American Journal of Epidemiology 172, 107–115) e in Stuart et al. (2011, *Using propensity scores to assess the generalizability of results from randomized trials*, Journal of the Royal Statistical Society Series A 174, 369–386); la connessione con la teoria formale di Pearl è sviluppata in Pearl e Bareinboim (2014, *External validity: from do-calculus to transportability across populations*, Statistical Science 29, 579–595).

Identificabilità non lineare con supervisione debole. Il modello iVAE (Identifiable Variational Autoencoder) è introdotto in Khemakhem, Kingma, Monti e Hyvärinen (2020, *Variational autoencoders and nonlinear ICA: a unifying framework*, AISTATS 2020, 2207–2217); la dimostrazione del teorema di identificabilità a meno di permutazioni e riscalamenti è in quel lavoro. Hyvärinen e Morioka (2016, *Unsupervised feature extraction by time-contrastive learning and nonlinear ICA*, NeurIPS 2016) e Hyvärinen, Sasaki e Turner (2019, *Nonlinear ICA using auxiliary variables and generalized contrastive learning*, AISTATS 2019) sviluppano il framework dell'ICA non lineare identificabile. Lippe, Magliacane, Löwe, Asano, Cohen e Gavves (2022, *CITRIS: Causal identifiability from temporal intervened sequences*, ICML 2022) estendono l'identificabilità iVAE al caso delle sequenze temporali con interventi. Ahuja, Hartford e Bengio (2022, *Weakly supervised representation learning with sparse perturbations*, NeurIPS 2022) dimostrano l'identificabilità con supervisione debole basata su perturbazioni sparse.

Causal GPT e integrazione strutturale nei modelli linguistici. Kıcıman, Ness, Sharma e Liu (2023, *Causal reasoning and large language models: opening a new frontier for causality*, arXiv:2305.00050) analizzano sistematicamente le capacità e i limiti dei modelli linguistici di grandi dimensioni nel ragionamento causale; la distinzione tra capacità correlazionale e capacità genuinamente causale è il tema centrale del lavoro. Zevcik, Pecka e Strakova (2023, *Can large language models reason about causal structures?*, arXiv:2309.01924) valutano empiricamente la performance su benchmark causali strutturati con struttura causale esplicita. Willig, Zečević, Dhami e Kersting (2022, *Causal inference with differentiable operator*, NeurIPS 2022) propongono un'architettura per l'integrazione del do-calculus come operatore differenziabile nei modelli neurali. L'analisi teorica delle limitazioni strutturali dei LLM come macchine associative di livello-1 è in Ortega et al. (2021, *Shaking the foundations: delusions in sequence models for interaction and control*, DeepMind Technical Report) e in Marcus (2018, *Deep learning: a critical appraisal*, arXiv:1801.00631).

Il principio di causa comune di Reichenbach. Il principio di causa comune (Common Cause Principle) è enunciato in Reichenbach (1956, *The Direction of Time*, University of California Press, capitolo IV). La sua connessione con la struttura di Markov causale e la d-separazione è discussa in Spirtes, Glymour e Scheines (1993, capitolo 2) e in Sober e Steel (2014, *Time and the symmetry of causal influence*, The British Journal for the Philosophy of Science 65, 821–848). La formulazione quantistica del principio di causa comune e il suo fallimento per correlazioni EPR è discussa in Wood e Spekkens (2015, *The lesson of causal discovery algorithms for quantum correlations: causal explanations of Bell-inequality violations require fine-tuning*, New Journal of Physics 17, 033002) — un'applicazione della teoria causale alla fondazione della meccanica quantistica che illustra la generalità del formalismo dei MAG: le correlazioni quantistiche violano il principio di causa comune per qualsiasi DAG classico, ma possono essere rappresentate da MAG con archi bidirezionali che codificano l'entanglement come "confondimento" senza interpretazione classica.